



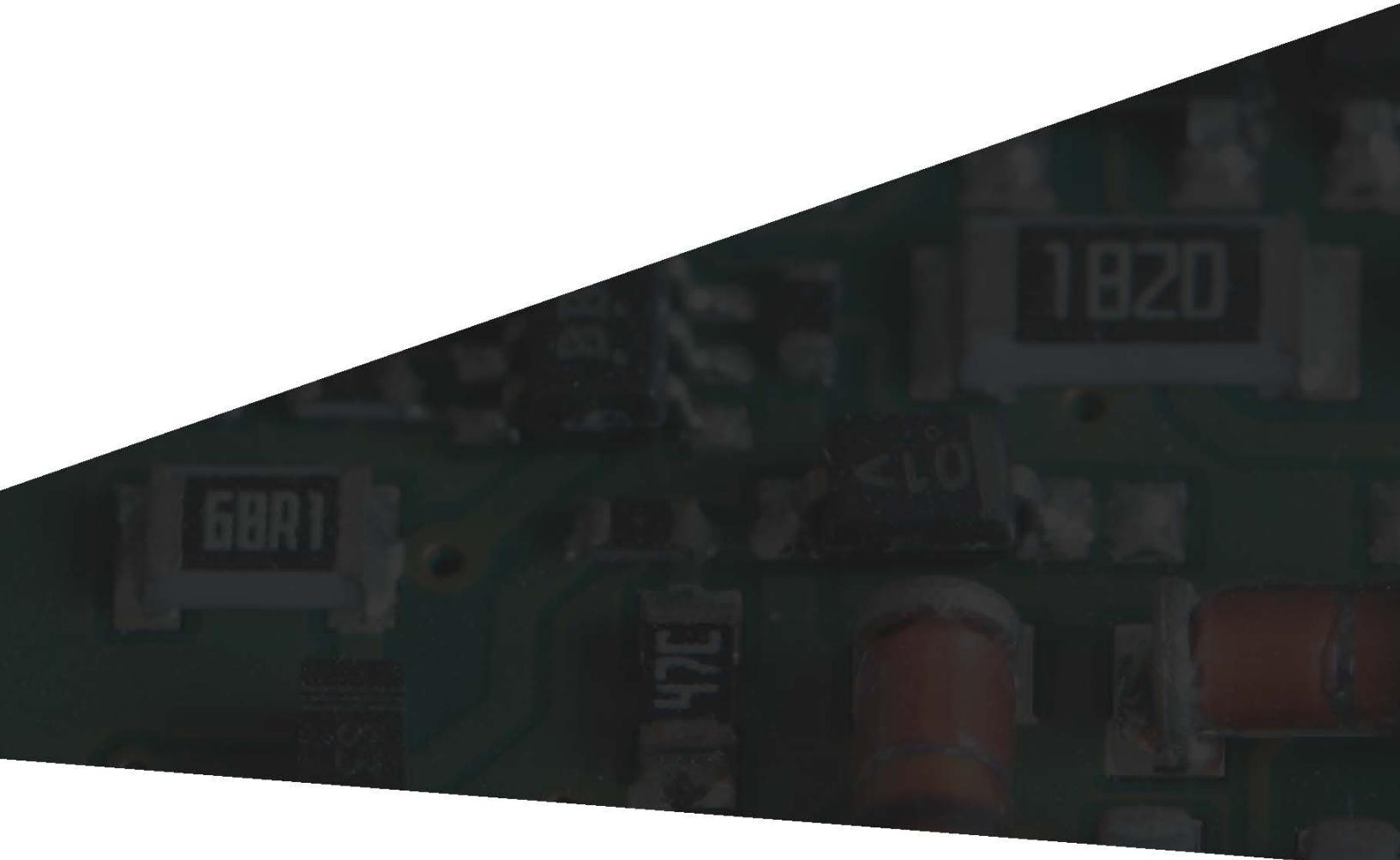
UNIVERSITY OF TORONTO
FACULTY OF INFORMATION



PRES 2012

Proceedings of the 9th International Conference on
Preservation of Digital Objects

October 1 - 5, 2012





PRES 2012

**Proceedings of the 9th International Conference on
Preservation of Digital Objects**

October 1 - 5, 2012

Editors: Reagan Moore, Kevin Ashley, Seamus Ross



UNIVERSITY OF TORONTO
FACULTY OF INFORMATION

Conference Organization

Program Committee

Reagan Moore

Chair

Reinhard Altenhoener	Nancy McGovern
Bjarne Andersen	Andrew McHugh
Andreas Aschenbrenner	Carlo Meghini
Kevin Ashley	Salvatore Mele
Tom Baker	Ethan Miller
Christoph Becker	David Minor
Karim Boughida	Jacob Nadal
Gerhard Budin	Heike Neuroth
Priscilla Caplan	Quyen Nguyen
Panos Constantopoulos	Achim Osswald
Paul Conway	Christos Papatheodorou
Robin Dale	Arthur Pasquinelli
Angela Dappert	David Pearson
Joy Davidson	Andreas Rauber
Michael Day	Seamus Ross
Janet Delve	Raivo Ruusalepp
Angela Di Iorio	Lisa Schiff
Jon Dunn	Michael Seadle
Miguel Ferreira	Robert Sharpe
Ellen Geisriegler	Barbara Sierman
David Giaretta	Tobias Steinke
Andrea Goethals	Randy Stern
Emily Gore	Stephan Strodl
Mariella Guercio	Shigeo Sugimoto
Mark Guttenbrunner	David Tarrant
Carolyn Hank	Manfred Thaller
Ross Harvey	Susan Thomas
Matthias Hemmje	Emma Tonkin
Leslie Johnston	Raymond van Diessen
Max Kaiser	Richard Wright
Ross King	Kate Zwaard
Amy Kirchhoff	
Hannes Kulovits	
Cal Lee	
William Lefurgy	
Jens Ludwig	
Maurizio Lunghi	

Chairs

Seamus Ross
Conference Co-chair

Kevin Ashley
Conference Co-chair

Angela Dappert
Workshop Co-chair

Carolyn Hank
Tutorial Co-chair

Cal Lee
Workshop Co-chair

Reagan Moore
Program Committee Chair

Raivo Ruusalepp
Tutorial Co-chair

Local Organizing Committee

Andrew Drummond
Chair

Katherine Shyjak

Kathleen Scheaffer

Ivan Sestak

Seamus Ross

Preface to iPRES 2012 Conference Proceedings

From October 1-5, 2012, the University of Toronto's Faculty of Information was pleased to host the ninth annual iPRES Conference. Previous conferences were held in Beijing (2004, 2007), Göttingen (2005), Ithaca, NY (2006), London (2008), San Francisco (2009), Vienna (2010), and Singapore (2011). The next conferences were planned for Lisbon (2013), Melbourne (2014), and Chapel Hill (2015).

The Organizing Committee was pleased to note that the event continued to garner significant interest, with well over 100 submissions received from 25 countries around the world. Most proposals came from the United States and the United Kingdom, but Portugal, Austria, Germany, Canada and the Netherlands were significant sources of proposals as well. Four workshops and five tutorial sessions were approved, as well as 42 papers and two panel presentations delivered during 16 sessions.

The conference hosted three keynote presentations: Steve Knight of the National Library of New Zealand gave a paper on "Implementing Guidelines for Preservation of Digital Heritage"; Kevin Ashley, Director of the UK's Digital Curation Centre on "Good Research, Good Data, Good Value: the Digital Curation Centre and the Changing Curation Landscape"; and Yunhyong Kim of blogforever, whose paper was entitled "Digital Preservation: A Game of Prediction". Technical sessions at the conference were on central preservation topics like Preservation Assessment, Training, Preserving Text Objects, Site Reports, Business Processes, Preservation Environments, Models, Concepts, and Community Approaches.

The conference also hosted an exciting poster/demo session that showcased the excellent work of some colleagues; presentations by students seemed especially impressive. The Poster Award went to Jamin Koo and Carol Chou for their presentation entitled "PDF to PDF/A: Evaluation of Converter Software for Implementation in Digital Repository Workflow". The poster/demo session, along with the conference banquet that followed, proved to be an excellent opportunity for academics, students, industry representatives and other professionals involved in digital preservation to network and share information.

Two corporate sponsors generously assisted the work of iPRES 2012: ExLibris Rosetta and Tessella-Preservica both provided time and resources to the Conference, and deserve considerable credit for their efforts in the field; the University of Toronto's Faculty of Information provided not only staff support to the conference, but also funded the first annual poster award.

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

Reagan Moore, Program Committee Chair
Kevin Ashley, Conference Co-Chair
Seamus Ross, Conference Co-Chair

TABLE OF CONTENTS

PRESERVATION IS KNOWLEDGE: A COMMUNITY-DRIVEN PRESERVATION APPROACH	
<i>SOPHIE DERROT, LOUISE FAUDUET, CLÉMENT OURY AND SÉBASTIEN PEYRARD.....</i>	1
DEVELOPING A COMMUNITY CAPABILITY MODEL FRAMEWORK FOR DATA-INTENSIVE RESEARCH	
<i>LIZ LYON, ALEX BALL, MONICA DUKE AND MICHAEL DAY.....</i>	9
CRISP: CROWDSOURCING REPRESENTATION INFORMATION TO SUPPORT PRESERVATION	
<i>MAUREEN PENNOCK, ANDREW N. JACKSON AND PAUL WHEATLEY</i>	17
AN ONTOLOGY-BASED MODEL FOR PRESERVATION WORKFLOWS	
<i>MICHALIS MIKELAKIS AND CHRISTOS PAPATHEODOROU.....</i>	21
INTEROPERABILITY FRAMEWORK FOR PERSISTENT IDENTIFIERS SYSTEMS	
<i>MAURIZIO LUNghi, EMANUELE BELLINI, CHIARA CIRINNÀ, BARBARA BAZZANELLA, PAOLO BOUQUET, DAVID GIARETTA AND RENÉ VAN HORIK</i>	29
CONVERSION AND EMULATION-AWARE DEPENDENCY REASONING FOR CURATION SERVICES	
<i>YANNIS TZITZIKAS, YANNIS MARKETAKIS AND YANNIS KARGAKIS.....</i>	38
CURATING THE SPECIFICITY OF METADATA WHILE WORLD MODELS EVOLVE	
<i>YANNIS TZITZIKAS, ANASTASIA ANALYTI AND MARY KAMPOURAKI</i>	46
PACKAGE FORMATS FOR PRESERVED DIGITAL MATERIAL	
<i>ELD ZIERAU.....</i>	54
RETHINKING AUTHENTICITY IN DIGITAL ART PRESERVATION	
<i>PERLA INNOCENTI</i>	62
DESCRIBING DIGITAL OBJECT ENVIRONMENTS IN PREMIS	
<i>ANGELA DAPPERT, SÉBASTIEN PEYRARD, JANET DELVE AND CAROL CHOU</i>	68
LDS3: APPLYING DIGITAL PRESERVATION PRINCIPALS TO LINKED DATA SYSTEMS	
<i>DAVID TARRANT AND LESLIE CARR</i>	76
AN ARCHITECTURAL OVERVIEW OF THE SCAPE PRESERVATION PLATFORM	
<i>RAINER SCHMIDT.....</i>	84
TOWARDS A LONG-TERM PRESERVATION INFRASTRUCTURE FOR EARTH SCIENCE DATA	
<i>ARIF SHAON, ESTHER CONWAY, BRIAN MATTHEWS, FULVIO MARELLI, UGO DI GIAMMATTEO, YANNIS MARKETAKIS, YANNIS TZITZIKAS, RAFFAELE GUARINO, HOLGER BROCKS AND FELIX ENGEL</i>	88
MIGRATION AT SCALE: A CASE STUDY	
<i>SHEILA MORRISSEY, VINAY CHERUKU, JOHN MEYER, MATTHEW STOEFLER, WILLIAM HOWARD AND SURESH KADIRVEL</i>	96
MANAGING MULTIDISCIPLINARY RESEARCH DATA : EXTENDING DSPACE TO ENABLE LONG-TERM PRESERVATION OF TABULAR DATASETS	
<i>JOÃO ROCHA DA SILVA, CRISTINA RIBEIRO AND JOÃO CORREIA LOPES.....</i>	105
ON THE APPLICABILITY OF WORKFLOW MANAGEMENT SYSTEMS FOR THE PRESERVATION OF BUSINESS PROCESSES	
<i>STEFAN PROELL, RUDOLF MAYER AND ANDREAS RAUBER</i>	109
DIGITAL PRESERVATION OF BUSINESS PROCESSES WITH TIMBUS ARCHITECTURE	
<i>MYKOŁA GALUSHKA, PHILIP TAYLOR, WASIF GILANI, JOHN THOMSON, STEPHAN STRODL AND ALEXANDER NEUMANN.....</i>	117

TOWARDS A DECISION SUPPORT ARCHITECTURE FOR DIGITAL PRESERVATION OF BUSINESS PROCESSES	
<i>MARTIN ALEXANDER NEUMANN, HOSSEIN MIRI, JOHN THOMSON, GONCALO ANTUNES, RUDOLF MAYER AND MICHAEL BEIGL</i>	125
AN OVERVIEW OF DIGITAL PRESERVATION CONSIDERATIONS FOR PRODUCTION OF "PRESERVABLE" E-RECORDS: AN INDIAN E-GOVERNMENT CASE STUDY	
<i>DINESH KATRE</i>	133
DEVELOPING RESEARCH DATA MANAGEMENT CAPABILITY: THE VIEW FROM A NATIONAL SUPPORT SERVICE	
<i>SARAH JONES, GRAHAM PRYOR AND ANGUS WHYTE</i>	141
ADVANCING DATA INTEGRITY IN A DIGITAL PRESERVATION ARCHIVE - EX LIBRIS AND THE CHURCH OF JESUS CHRIST OF LATTER-DAY SAINTS	
<i>GARY WRIGHT AND NIR SHERWINTER</i>	149
FORMATS OVER TIME: EXPLORING UK WEB HISTORY	
<i>ANDREW N. JACKSON</i>	154
FROM CATALOGUING TO DIGITAL CURATION: THE ROLE OF LIBRARIES IN DATA EXCHANGE	
<i>SUSAN K. REILLY</i>	158
THE COMMUNITY-DRIVEN EVOLUTION OF THE ARCHIVEMATICA PROJECT	
<i>COURTNEY MUMMA AND PETER VAN GARDEREN</i>	163
AUTHENTICITY MANAGEMENT IN LONG TERM DIGITAL PRESERVATION OF MEDICAL RECORDS	
<i>SILVIO SALZA AND MARIELLA GUERCIO</i>	171
FUTURE-PROOF PRESERVATION OF COMPLEX SOFTWARE ENVIRONMENTS	
<i>KLAUS RECHERT, ISGANDAR VALIZADA AND DIRK VON SUCHODOLETZ</i>	179
PRACTICAL FLOPPY DISK RECOVERY STUDY - DIGITAL ARCHEOLOGY ON BTOS/CTOS FORMATTED MEDIA	
<i>DIRK VON SUCHODOLETZ, EUAN COCHRANE, DAVID SCHMIDT AND RICHARD SCHNEIDER</i>	183
DUPLICATE DETECTION FOR QUALITY ASSURANCE OF DOCUMENT IMAGE COLLECTIONS	
<i>REINHOLD HUBER-MÖRK, ALEXANDER SCHINDLER AND SVEN SCHLARB</i>	187
AUDIO QUALITY ASSURANCE: AN APPLICATION OF CROSS CORRELATION	
<i>JESPER SINDAHL NIELSEN AND BOLETTE AMMITZBØLL JURIK</i>	195
EVALUATING AN EMULATION ENVIRONMENT: AUTOMATION AND SIGNIFICANT KEY CHARACTERISTICS	
<i>MARK GUTTENBRUNNER AND ANDREAS RAUBER</i>	201
DIGITAL PRESERVATION OF NEWSPAPERS: FINDINGS OF THE CHRONICLES IN PRESERVATION PROJECT	
<i>KATHERINE SKINNER, MARTIN HALBERT, MATT SCHULTZ AND MARK PHILLIPS</i>	209
BLOGS AS OBJECTS OF PRESERVATION: ADVANCING THE DISCUSSION ON SIGNIFICANT PROPERTIES	
<i>KAREN STEPANYAN, GEORGE GKOTSISS, HENDRIK KALB, YUNHYONG KIM, ALEXANDRA I. CRISTEA, MIKE JOY, MATTHIAS TRIER AND SEAMUS ROSS</i>	218
CHALLENGES IN ACCESSING INFORMATION IN DIGITIZED 19TH-CENTURY CZECH TEXTS	
<i>KAREL KUCERA AND MARTIN STLUKA</i>	225
ESA USE CASES IN LONG TERM DATA PRESERVATION	
<i>MIRKO ALBANI, ROSEMARIE LEONE AND CALOGERA TONA</i>	229

REQUIREMENTS ELICITATION FOR A LONG TERM DIGITAL PRESERVATION SYSTEM: A CASE STUDY FROM THE FINANCIAL SECTOR	
<i>Claudia-Melania Chituc and Petra Ristau</i>	236
WEB ARCHIVING EFFORT IN NATIONAL LIBRARY OF CHINA	
<i>Yunpeng Qu</i>	244
ADDRESSING DATA MANAGEMENT TRAINING NEEDS: A PRACTICE-BASED APPROACH FROM THE UK	
<i>Laura Molloy, Simon Hodson, Stéphane Goldstein and Joy Davidson</i>	248
AHEAD OF THE CURV: DIGITAL CURATOR VOCATIONAL EDUCATION	
<i>Laura Molloy and Ann Gow</i>	256
PRESERVING ELECTRONIC THESES AND DISSERTATIONS: FINDINGS OF THE LIFECYCLE MANAGEMENT FOR ETDs PROJECT	
<i>Katherine Skinner, Martin Halbert and Matt Schultz</i>	261
PRESERVATION WATCH: WHAT TO MONITOR AND HOW	
<i>Christoph Becker, Kresimir Duretec, Petar Petrov, Luis Faria, Miguel Ferreira and Jose Carlos Ramalho</i>	266
ASSESSING DIGITAL PRESERVATION CAPABILITIES USING A CHECKLIST ASSESSMENT METHOD	
<i>Gonçalo Antunes, Diogo Proença, José Barateiro, Ricardo Vieira, Jose Borbinha and Christoph Becker</i>	274
EVALUATING ASSISTED EMULATION FOR LEGACY EXECUTABLES	
<i>Setha Toshnial, Geoffrey Brown, Kevin Cornelius, Gavin Whelan and Enrique Areyan</i>	282
AUTOMATED DIGITAL PROCESSING AT THE BENTLEY HISTORICAL LIBRARY	
<i>Nancy Deromedi, Michael Shallcross</i>	290
AGGREGATING A KNOWLEDGE BASE OF FILE FORMATS FROM LINKED OPEN DATA	
<i>Roman Graf, Sergiu Gordea</i>	292
BIBLIOBLOGGERS' PRESERVATION PERCEPTIONS, PREFERENCES, AND PRACTICES	
<i>Carolyn Hank, Cassidy R. Sugimoto</i>	294
POSTER 'PRESERVATION THROUGH ACCESS: THE AHDS PERFORMING ARTS COLLECTIONS IN ECLAP AND EUROPEANA'	
<i>Perla Innocenti, John Richards</i>	296
A DIGITAL REPOSITORY YEAR: ONE MUSEUM'S QUEST FOR THE BASICS	
<i>Paula Jabloner, Katherine Kott</i>	299
PDF TO PDF/A CONVERTER EVALUATION	
<i>Carol Chou, Jamin Koo</i>	301
ON THE COMPLEXITY OF PROCESS PRESERVATION: A CASE STUDY ON AN E-SCIENCE EXPERIMENT	
<i>Rudolf Mayer, Stephan Strodl, Andreas Rauber</i>	304
THE COMMUNITY-DRIVEN EVOLUTION OF THE ARCHIVEMATICA PROJECT	
<i>Peter Van Garderen, Courtney C. Mumma</i>	306
PRESERVING ELECTRONIC THESES AND DISSERTATIONS: FINDINGS OF THE LIFECYCLE MANAGEMENT FOR ETDs PROJECT	
<i>Martin Halbert, Katherine Skinner, Matt Schultz</i>	314
CREATING VISUALIZATIONS OF DIGITAL COLLECTIONS WITH VIEWSHARE	
<i>Trevor Owens, Abigail Potter</i>	319

SCALABLE CONTENT PROFILING FOR PRESERVATION ANALYSIS	
<i>PETAR PETROV, CHRISTOPH BECKER</i>	322
DEFINING DIGITAL CURATION THROUGH AN INTERACTIVE, INFORMAL CRITICAL DELPHI APPROACH	
<i>LORI PODOLSKY NORDLAND, CAROLYN HANK</i>	324
BWFLA – PRACTICAL APPROACH TO FUNCTIONAL ACCESS STRATEGIES	
<i>KLAUS RECHERT, DIRK VON SUCHODOLETZ, ISGANDER VALIZADA</i>	327
WILL FORMAL PRESERVATION MODELS REQUIRE RELATIVE IDENTITY?	
<i>SIMONE SACCHI, KAREN M. WICKETT, ALLEN H. RENEAR</i>	329
TRAINING NEEDS IN DIGITAL PRESERVATION A DIGCURV SURVEY	
<i>CLAUDIA ENGELHARDT, STEFAN STRATHMANN</i>	331
RETROCOMPUTING AS PRESERVATION	
<i>YURI TAKHTEYEV, QUINN DUPONT</i>	334
DURA CLOUD, CHRONOPOLIS AND SDSC CLOUD INTEGRATION	
<i>ANDREW WOODS, BILL BRANAN, DAVID MINOR, DON SUTTON, MICHAEL BUREK</i>	337
DEMO – AN INTEGRATED SYSTEM-PRESERVATION WORKFLOW	
<i>ISGANDAR VALIZADA, KLAUS RECHERT, DIRK VON SUCHODOLETZ, SEBASTIAN SCHMEIZER</i>	340

Preservation Is Knowledge: A community-driven preservation approach

Sophie Derrot

Department of
Legal Deposit
sophie.derrot@bnf.fr

Louise Fauduet

Department of
Preservation and
Conservation
louise.fauduet@bnf.fr

Clément Oury

Department of
Legal Deposit
clement.oury@bnf.fr

Sébastien Peyrand

Department of
Bibliographic and Digital
Information
sebastien.peyrand@bnf.fr

Bibliothèque nationale de France (BnF, National Library of France)

Quai François Mauriac

75706 Paris Cedex 13

ABSTRACT

In the beginning, SPAR, the National Library of France's repository, was designed as the OAIS softwarified. It was intended to be a "full OAIS", covering all preservation needs in one tidy system. Then as its potential revealed itself across the library, high hopes arose for a do-it-all digital curation tool. Yet in day to day preservation activities of the BnF, it turns out that SPAR's growth takes a practical approach to the essentials of preservation and the specific needs of communities. Renewed dialogue with producers and users has led to the addition of functions the digital preservation team would not have thought of. This is very clear in what has been created to ingest the BnF's web archives into SPAR, giving the community more information on their data, and in what is taking shape to deal with the BnF's administrative archives, adding new functionalities to the system. The difference between what preservations tools and what curation tools should be at the BnF will have to be examined over time, to ensure all the communities' needs are met while SPAR remains viable.

Keywords

Digital Curation; Preservation Repository; Web Legal Deposit; Digital Archives.

1. INTRODUCTION: BUILDING A REPOSITORY

In the beginning SPAR was designed as a comprehensive digital preservation tool. But we had to reduce its initial scope, and ended up using it for wider purposes than preservation.

1.1 The Original Vision

The National Library of France has been working on building a digital repository to preserve its assets since 2005. This project, called SPAR (Scalable Archiving and Preservation Repository), is intended to be as comprehensive a digital preservation tool as possible. Quite logically, it initially encompassed all the various aspects of digital preservation:

- **Full range of functions.** SPAR meant to implement all the OAIS entities that could be automated: ingest workflow through Ingest, Storage and Data Management functions; dissemination workflow through Storage, Data Management and Access functions; last but not least, a preservation workflow through Preservation Planning and Administration interfaced with the aforementioned workflows.

- **Full range of assets.** SPAR aimed at storing and preserving a very wide range of assets with heterogeneous legal statuses and technical characteristics, from digitized text, image, video and audio content to digital legal deposit, digital archival records and databases, and third-party archived content.
- **The range of preservation levels.** On this double workflow-and content-oriented approach, SPAR aimed at allowing all possible preservation strategies (bit level refreshment and media migration, format migration and emulation) depending on the legal and technical aspects of the corresponding asset.

1.2 Making It Feasible: Prioritizing the Developments and Tightening Up the Scope

This long-term vision could not be achieved in a fully-fledged system and organization in a single run, so the problem and vision had to be split into discrete, manageable, prioritizable bits. This resulted in two aspects:

- **1.2.1 Splitting the Functions: a Modular Approach**
SPAR was designed as a set of interrelated modules, which allowed the system to be developed and updated on a per-module basis. Each OAIS entity was fully implemented as an autonomous module in the system, which communicates with other modules through standard RESTful web services. But all functions did not have the same urgency: before assessing any preservation plans on objects, they first had to be ingested in, and accessed from, a repository. Thus, the development of the Preservation Planning module had to be delayed.

1.2.2 Segmenting the Document Sets: the Tracks and Channels

The preservation policies differed depending on the documents:

- **Legal aspects:** the digital assets to be preserved can be subject to various legal frameworks: legal deposit law; archival records preservation and curation duty law; intellectual property laws and their exceptions for heritage institutions; convention with third party organizations for third party archiving; donations; and so on. Depending on the legal framework of the assets, the library will not be allowed the same range of actions to preserve them.
- **Life cycle management issues:** sometimes it is crucial to have the ability to fully delete all the versions of an AIP in a repository for legal purposes (e.g. for archival records); sometimes it is the exact opposite, with a guarantee that no

deletion of any “version 0” will ever be done (e.g. for born-digital legal deposit); finally, in some cases this might change over time (e.g. digitization, depending on the condition, rarity and complexity of the source physical document);

- **Preservation strategy / Significant properties:** sometimes the content and layout must be preserved (e.g. digitized books), sometimes the top-level priority is the intellectual content (e.g. some archival records), sometimes the user experience is almost as important as the content itself (e.g. “active content” like video games, or born-digital heritage like web archives).

These assets could be grouped in different ways, but few were really satisfactory. Grouping them **by document category** was not very efficient, because different policies could be applied to the same kind of document depending on what is the National Library of France’s obligation to preserve it. For example, a born-digital asset will not necessarily be managed the same way if it has been ingested as Legal Deposit or submitted by a third party organization. Grouping the assets on the basis of the **curation services** responsible for them was deemed incompatible with long-term preservation as it would be based on the organization chart, which frequently changes over time. Finally, a **legal framework distinction** seemed well-suited but insufficient, since the same legal framework can be applied to objects with heterogeneous technical characteristics.

However, all these aspects were to be taken into consideration somehow. In other terms, the problem was to find the right balance between the legal, technical and organizational aspects.

This was achieved by grouping the assets into **tracks and channels**. Each track had a set of digital objects belonging to the same legal framework and overall curatorial characteristics, and targeted at a particular user community. Example of tracks included:

- Preservation of digitized books, periodicals and still images
- Audiovisual content
- Web legal deposit
- Negotiated legal deposit
- Archival records preservation
- Donations and acquisitions against payment

Each track is then subdivided into one or more channels, which group together assets with homogeneous technical characteristics.

The first track and channel to be developed was the digitization of books, periodicals and still images, for pragmatic reasons: achieving a critical mass of archived objects very quickly to secure preservation budgets; and achieving a good proportion of the metadata management needs by coping with the best known – and thus most documented – content.

1.3 Making It Real: Back to the Reality Principle

When developing the core functions of SPAR, the team quickly faced huge delays in developments, partly because of the “research and development” aspect of the project and the very specific needs of the BnF in terms of scale, performance and variety of data objects. The functional scope had thus to be reduced. This choice was made on the basis of two criteria:

- Where were the development challenges and failure risks highest?
- What could be abandoned, at least for the moment, while maintaining an up-and-running consistent workflow?

The Access functions were therefore abandoned, as both the most risky part and the dispensable one. For the digitization preservation track alone, the BnF’s needs in terms of AIP to DIP transformations (thumbnails, low and medium resolution for web browsing, PDF downloadable content, etc.) were very hard to scale up to the mass of collections at stake (1,5 million DIPs).

From the perspective of our aforementioned different repository workflows, the Ingest, Storage and Data Management modules had priority over the Access and Rights management ones. The library Information System already had existing, though perfectible, applications to manage the digital library and the rights management part. So the scope of our Access module was reduced to the mere dissemination of AIPs. The access and rights management functions were reported to the Access existing applications and Designated User communities for each track.

1.4 It's Alive! Making It Run and Keeping It Growing

With the aforementioned phasing methodology and scope reduction, SPAR went operational in May 2010 for its first core functions and track. From then on, the developments strongly focused on ingesting new content by working on new tracks and channels:

- **Third party storage** (summer 2010): functions to receive content from outside the library
- **Audiovisual track**: audio and video digitization, and CD-audio extraction (spring 2011): audio and video files analysis functions, and management of complex structures such as multimedia periodicals;
- **Web legal deposit** (spring 2012): management of container file analysis (especially ARC files; see below)

Advanced systems administration functions were also added during the first year, and they mostly consisted in helping the IT team manage workflows as efficiently as possible, e.g. to plan mass AIP dissemination and mass fixity checks.

In other terms, the development policy was centered around SPAR as digital library stacks: optimizing the ingest workflows, receiving new kinds of assets (and developing the functions required to do this). This resulted in an increased shared knowledge between curators and preservationists. For each new track, during the design stages, this was initiated with the exchange of knowledge about the digital preservation tool on one hand and the assets at stake and user community needs on the other hand. However, this knowledge of the preserved assets was unexpectedly increased by the preservation tool itself in action.

1.5 Using It: a Digital Collection Knowledge Utility?

The first concrete effect SPAR had on collection curation was indeed the increased available knowledge that was gained on the ingested digital assets, especially regarding their history and overall technical characteristics. The audiovisual track was a good example of such added knowledge, acquired during the tests:

- **Image compression problems:** the curators discovered that some CD boxes and phonogram image shots were LZW-compressed, a format considered risky at the BnF because there was no in-house expertise on it. These images had to be de-compressed before they could be ingested.

- **Unexpected video frame rate structure:** unorthodox 15 frames-GOPs (Group of Pictures)¹ and even variable ones were found. As the content could all the same be displayed, it was decided to ingest and preserve them “as is” but keep all these characteristics in the repository metadata where they could be tracked down.

These two facts were unknown to the library’s audiovisual content curators, since they had no impact on the rendering. In this way SPAR’s file analysis functions² allowed increased knowledge of the collection’s technical characteristics. From a long-term perspective, it lowered preservation risks by removing some risky features (e.g. compression) or documenting them (e.g. the GOP) so that the corresponding files could be specifically retrieved in the future.

These features were made possible by SPAR’s data management module, which documents nearly all the information required for our AIPs (technical characteristics and file formats, operations performed from creation to the present, policies for ingest and preservation, structure and basic description of the intellectual content) in the form of a RDF database accessible through a SPARQL endpoint [5].

In the end, the design and testing was a very special moment where curators found SPAR gave them a better grasp of the nature and arrangement of their collections. This demonstrated one particular benefit of SPAR where the primary aim was not preservation but rather knowledge of the assets, and therefore curation. This aspect gained even more momentum in the web archives track and the digital archives track.

2. WEB ARCHIVES

2.1 A Track with Very Specific Needs

Since 2006, thanks to an extension of its mission of legal deposit, BnF is mandated to collect and preserve the French publications online [6]. The whole set of data publicly available on the French Internet is concerned: videos, public accounts on social networks, blogs, institutional websites, scientific publications, and so on. BnF uses robots (crawlers) that harvest data from the web and store it in ARC files³. The major characteristics that guided the development of the web archives track in SPAR were determined by the specific legal and technical status of these collections:

- legally: long-term preservation, forbidding the deletion of the data, the obligation of preserving the original documents as collected and, at the same time, to give access to the data ;
- technically: data which result from an automatic crawl and even from a succession of different production workflows (by the BnF but also by others partners, by different crawlers, etc.), a wide range of formats and objects.

¹ The Group of Pictures is a way to document how the moving image stream is divided into full frames and, if any, intermediary frames that only list the differences from the next frame in a predictive fashion. See http://en.wikipedia.org/wiki/Group_of_pictures.

² SPAR identifies formats with a Java-packaged File UNIX command, and analyses image and text with JHOVE, audio and video with Mediainfo, and ARC container files with JHOVE2.

³ ARC is a container format designed for web archives (see <http://archive.org/web/researcher/ArcFileFormat.php>). Its evolution, the WARC format, is an ISO standard (28500:2009)

Of course, the digital legal deposit track’s design benefited from the development and reflections on the pre-existing tracks (audiovisual and digitization tracks), and will in turn nourish the next ones (third-party, negotiated legal deposit and administrative tracks). For example, as opposed to the previous tracks, the legal deposit one was bound to strictly forbid the modification or deletion of the original data objects: what the BnF collects by legal deposit must be kept and preserved for access. This question also concerns the administrative archive (see below).

Another example is the preservation of the user experience. For the web archive, not only the content itself, but also its environment of consultation matters; this is not the case for the digitization preservation track for books, periodicals and still images, where content is predominant. To this end, the crawler declares itself as a browser; in order to ensure the harvesting of the content as it was offered to the user. The access to the archive is by an embedded browser and the data must be collected and preserved to enable it to be displayed as on the live web.

2.2 The Challenge of Diversity

It is planned for the web archives to enter SPAR in the automatic legal deposit track. In a way, this track is probably the one which is the most deeply linked with the basic aims of SPAR. The obligation of long-term preservation is impossible under the current conditions of storage of the collections (hard drives and storage bays with no preservation system), and SPAR is the only way for the Library to fully perform its duty. In addition, the diversity of these collections increases the difficulty of preserving and knowing them; only a system dedicated to the treatment of digital collections could permit us to curate such objects.

During the implementation of this track, solutions to several technical challenges had to be found. One of the main issues for web archives preservation is the lack of information on harvested file formats: the only available one is the MIME type sent by the server, which is frequently wrong [7]. To this end, the developments included the design of a Jhove2 module for the ARC format⁴. It is able to identify and characterize ARC files but also the format of the files contained within them. This tool will bring the librarians unprecedented knowledge on their collections. Along the same lines the “containerMD” metadata scheme⁵ was implemented to allow the recording of technical information for container files.

BnF web archive collections are made of several data sets which came from different harvesting workflows [8], in different institutions with various practices (the BnF, the Internet Archive foundation, Alexa Internet which worked with IA). SPAR was a natural choice for preserving these web archives, but some adjustments were necessary on both sides, and particularly the homogenization of the different collections into one data model. Inside the track, five channels were distinguished, according to the workflow using for the harvest. Not every channel has the same level of description and metadata. The librarians knew from the beginning the major differences between the channels, but this knowledge was markedly improved by the implementation of the track and the necessary work of homogenization.

⁴ See <https://bitbucket.org/jhove2/main/wiki/Home>. Development of a WARC module for Jhove2 is currently performed by the Danish Netarchive.dk team.

⁵ On containerMD, see <http://bibnum.bnf.fr/containerMD>.

2.3 Knowing Collections by Implementation

The SPAR team is now close to the end of the implementation of the digital legal deposit track, which began two years ago. This provides an opportunity to consider the choices made at the beginning of this work.

RDF was chosen as the indexation model in SPAR. The triple-store capacity is limited, and the stand was taken not to index some data of the ARC files, especially the associated files. During a crawl performed by Heritrix and NAS, files are produced with reports and metadata about the crawl (crawl log, hosts reports, seed list); the large size of these files made their complete indexation impossible. Thus it is impossible to obtain by a SPARQL query the list of the harvest instances containing a certain domain name. This was a conscious choice made during the development of the track, and therefore a known limit of the knowledge about the collections.

On the other hand, a lot of metadata are indexed and therefore can support a SPARQL query. Especially, SPAR ingests reference information about agents performing preservation operations, which can be performed by humans (administrators, preservation experts), software tools (identification, characterization and validation tools) and processes in SPAR (such as the ingest and package update process). Performing these requests allows precious statistic, technical or documentary information to be retrieved about the collections:

- for example, the list of the crawlers (“agent”) and the version used by channel can be produced by querying the agent linked to the harvest event with a role of “performer”:

Table 1. Response to a SPARQL query on crawling software tools for each channel

channelId	agentName
fil_dl_auto_cac	Heritrix 1.10.1
fil_dl_auto_cac	Heritrix 1.12.1
fil_dl_auto_cac	Heritrix 1.14.0
fil_dl_auto_cac	Heritrix 1.14.2
fil_dl_auto_cia	Heritrix 1.14.1
fil_dl_auto_cia	Internet Archive
fil_dl_auto_his	Alexa Internet
fil_dl_auto_htt	HTTTrack 3.10
fil_dl_auto_htt	Alexa Internet
fil_dl_auto_htt	HTTTrack 3.30
fil_dl_auto_nas	Heritrix 1.14.3
fil_dl_auto_nas	Heritrix 1.14.4

- another example is the list of harvest instances with “elections” in their title or description:

Table 2. Response to a SPARQL query on harvest instances concerned by the electoral crawls

Harvest definition	Title
ark:/12148/bc6p03x7j.version0.release0	BnF elections 2002
ark:/12148/bc6p03z7s.version0.release0	BnF elections 2004
ark:/12148/bc6p03zd5.version0.release0	BnF elections 2007

At the end of the implementation process, testing the possibilities of SPARQL queries on this track allowed the discovery of a few bugs or mistakes. But most of all, it gave the opportunity to fully consider the tool offered for the management of the collections.

The heterogeneity of data models between web archives from different periods was a strong obstacle that prevented from having

a common view on the BnF collections. The alignment of those data models and the possibility of requesting all collections the same way thanks to the data management module will permit getting similar metrics for all kind of assets. In that way SPAR will help providing the BnF the statistics and quality indicators necessary to measure and evaluate its collection. A list of these indicators is currently designed by a dedicated ISO working group, whose draft recommendations influenced the implementation of the web archives track⁶.

Testing the preingest phase for the test dataset also allowed the application of comprehensiveness tests. Each ARC metadata AIP contains a list of all ARC files produced by the harvest instance, as the outcome of a harvest event. Automatically comparing such lists with the ARC data files actually ingested in SPAR may prove very useful with old collections, for which there is a risk of losing data. It ensures too that incomplete or defective datasets cannot enter SPAR, which could otherwise be problematic for the preservation process. This new feature has been added to the administration module GUI.

2.4 Outside of SPAR

SPAR is the natural way to preserve the web archives over the long term. But in the meantime, several migration and packaging operations are performed outside of SPAR, which could have been thought of as typical preservation operations. For example, the BnF is planning to migrate all its ARC files to WARC files, thanks to specific migration tools. These tools will not be part of the SPAR workflow, but will be external. However, all the operations on the collections will be documented in the system, as the PREMIS data model, the cornerstone for SPAR’s RDF data model, allows the monitoring of each “Event” related to a file or a file group. The traceability of this kind of operation is key information to the curation of digital collections.

On the later crawls, the data harvested by the Heritrix are repackaged and enriched by metadata on the harvest by the curator tool, NAS. So the majority of the metadata on the harvest itself is pre-existing and therefore quite easily controlled by the librarians. This could be seen as easier on a daily basis, but it is also restrictive because every modification of the external tool must be made in the perspective of the ingest in SPAR. It forces the librarians to consider their collections from a preservation point of view and reinforce the consistency of the collection.

3. A DIFFERENT KIND OF COMMUNITY: ARCHIVES IN THE LIBRARY

3.1 Yet Another Track

During 2012, the SPAR team has been focusing on the ingestion of archives. The plan is to build experience with the BnF’s own documents, with a view to expanding its third-party preservation offer in the process, to records and archives in other institutions. In preparing the requirements for a new tender to further develop the system, starting this fall, the preservation team is learning yet again how taking into account new producers and designated communities is pushing the services of the Archive, and even its philosophy, in new directions.

⁶ The ISO TC46/SC8/WG9 is currently working on a Technical Report (ISO TR 14873) on Statistics and Quality Issues for Web Archiving that will be validated and published within a year. See also [2] on the question of web archive metrics.

3.1.1 Different Legal Requirements

Although France has promulgated a unified code of law for its cultural heritage, the *Code du Patrimoine*⁷, in 2004, it does not imply that a library could pick up archives and know what to do with them. And yet, the BnF has been producing records of its activities, and has been managing its own administrative archives, from the paper ages to the digital times. It has created a dedicated bureau to do so, recruiting archivists trained in the specificities of records management and the curation of historical archives, regardless of their medium.

Thus, in order to preserve the growing digital part of these archives, the SPAR team is now dealing with a new kind of producer and user community, and information managed under different rules of law. In the system, this translates into the creation of a new "track" for "administrative and technical production".

The main constraints that differ widely from the digital preservation team's previous endeavors with digitization and legal deposit stem from the added complexity of the information lifecycle: there is a much higher chance that information may be accessed and reused to create new versions of documents, and, above all, it may, and sometimes must, be deleted. The law on public archives requires that, once they are no longer in active use, documents that are not important for administrative, scientific, statistical or historical purposes should be weeded out of archives. Should different service levels then be applied to different stages in the lifecycle? Up to which point can sorting and eliminating records be automated? The role of SPAR in this process is beginning to take form.

3.1.2 A Specific Technical Environment

While acclimating to this different legal context, the digital preservation team also has to take into account an increased variety of documents and data, and specific work environments. The BnF's archives encompass the usual office documents — word processing, spreadsheets, slides and PDFs, — as well as a long trail of varied file formats, and the number of documents not in a format from the Microsoft Office suite increases steadily over the years. The library also produces highly specific records of its activities using specialized business software, such as financial databases or architectural plans.

From the first overview of this "track" in SPAR, it had thus been posited that several separate "channels" would be required to deal with the various types of records from the library's activities, and interact with their different production environments. A choice was made to focus this year on what is supposed to be the most standard of those channels, the one for regular office work records.

Yet there are challenges, given that the documents are stored and classified using proprietary software, IBM Lotus Notes. In addition, the BnF's agents tend to use this software in an idiosyncratic manner, in spite of the library archivists' efforts over the past years to fit it closely to the library's records production. Moreover, it would seem that the designated community for this part of the Archive is the largest SPAR has ever had to serve so far: producers and users of the administrative records are the library agents as a whole.

⁷ The latest version of which is available, in French, at <http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006074236> (accessed 21 May 2012).

Their representatives in the working group piloting SPAR's developments are brand new to the process, and bring a new and highly technical knowledge to the building of the repository: the BnF's two archivists have experience in records management and archival law, its Lotus administrator understands the workings of data and metadata in the document-oriented databases. Following the needs of the designated community for this new "channel for administrative production" is again changing the original contour of the SPAR system.

3.1.3 A New Set of Challenges

With the first tender for the development of SPAR's software ending in January 2012, it was decided that a first study of the requirements for the Administrative Channel would serve as an evaluation tool for potential new contractors. In the few months of this first investigation of the needs for the preservation of the BnF's administrative archives, issues emerged regarding which data and metadata to collect, create, preserve and disseminate. For instance, SPAR's team had never had to deal before with

- a greater attention to the issue of integrity and authenticity: the records and archives world is much more concerned with the possibility that a document may be required in a judicial context, where it will be necessary to prove that it has not been tampered with. What this means in a digital environment has yet to be clarified by jurisprudence;
- a lifecycle that may require documents to be accessed and modified in their original production environment, and, later on, in an updated or different business management environment that would have to interpret the original data and metadata correctly, and allow complex use of it;
- a more pressing need for a mechanism to delete AIPs and trace those deletions.

Other institutions and companies have had to solve such problems before⁸, but in the context of a library, and at this point in the development of SPAR, they are likely to be the source of a whole crop of new features in the system.

3.2 How to Manage: Verify, Migrate, Delete?

Given that preserving records is not necessarily new business, the BnF did not set out to reinvent the wheel, but existing solutions for records management and digital archiving did not fit the library's preservation plan:

- the core functions of SPAR have been designed to be generic, i.e. deal with information packages from all tracks and channels with the same processes. Introducing a whole new system was not considered an option;
- the requirements for the modernization of the French administration have first focused on a specific set of records that do not match the diversity of data in the BnF's Lotus Notes bases, nor its specific structure.

There is a national standard for the exchange of archival data ("Standard d'échange de données pour l'archivage", SEDA⁹) that the BnF will implement to deal with the messages and metadata attached to information transfer between producers, Archive and

⁸ Regarding rendering office documents for instance, Archives New Zealand's recent report is illuminating [4].

⁹ Schemas, tools and profiles are available, in French, at <http://www.archivesdefrance.culture.gouv.fr/seda/> (accessed 14 May 2012). A version 1.0 of the standard is in the works.

users. However, to create an interface between Lotus Notes and SPAR, this standard might not be fitting or necessary.

Moreover, the integrity of the BnF's Lotus databases is secured by multiple replications. The role of SPAR in the preservation of administrative production was rapidly defined by the working group as long term preservation of archives, not bit level preservation in the records management processes. Which of the records management processes, then, have to be maintained when the lifecycle of the records brings them to the point when they are ingested into the SPAR repository?

3.2.1 The Problem with Signatures

The BnF's archivists and IT specialists have secured authenticity in the library's records management through user authentication, digital signatures — to prove a record's origin, and access control lists — to manage access rights to the application, document, view and item levels. Whether this information can, and should, be carried over to the SPAR repository is a question the BnF has to research further. At this point in the specifications of the future Administrative Channel, it seems that it would be a Sisyphean task to renew the certificates associated with the signatures regularly since the certificates have a lifetime of a few years, and most of the BnF's archives reaching SPAR are to be preserved indefinitely.

It may however be useful to verify each document's signature at the moment the documents are transferred from the Lotus databases to the first stages of the ingest process. The signature files themselves might even be included in the METS manifest of the information packages if their durability can be proved. It seems likely, however, that the main assurance of the records' authenticity will come from sustaining and demonstrating the trustworthiness of SPAR's processes. This actually agrees with the practices of the producers and users of this Administrative Channel: the BnF's archivists rely as much on available documentation as on their skills in analyzing records for clues about their provenance, authenticity and integrity. In the working group, they told the preservation team they did not expect digital records to conform to an authenticity standard that has never been required in the paper world.

3.2.2 Conciliating Preservation and Access: Instant Migration

As can be expected in a large institution such as the BnF, constraints about number of users and budget, licensing fees in particular, make it difficult to switch to the latest and most easily preserved technologies. The library still relies on the 2003 Microsoft Office Suite, for example, with only binary formats available so far. Furthermore, the diversity of the library's activities means that no limit can be imposed on the file formats used, although the use of Word, Excel and PowerPoint files as attachments is facilitated, and represents about half of the files present in the databases.

The Administrative Channel processes must guarantee that the archived documents can be rendered again at any time in the Lotus Notes interface, in all their diversity. Which means that the specific structure of the Lotus document-oriented databases must be preserved as well: each document is stored in a series of fields, regardless of what could be considered data, or metadata. The items in a document encompass detailed provenance information, as well as rich content and attachments. Lotus provides an export and import function in a proprietary XML format, DXL, that may solve the issue.

Meanwhile, the service level for these documents in SPAR must be better than the bit-level preservation in an extraction in a proprietary XML format, and it must guarantee not only future rendering, but also modification of the data: relying on emulation alone might not be enough. The SPAR team is investigating the following approaches so far (see Figure 1):

- recording the visual aspect of the original document in a standardized format, using Lotus' PDF export capabilities for instance;
- taking the encapsulated files out of the DXL export of the document, making them easier to identify, characterize or migrate over time;
- transforming the remaining data in the DXL files to an open format, such as XHTML;
- making it all apparent in the "USE" attribute of the corresponding file groups in the METS manifest of the information packages.

Historically, files that are considered the focus of preservation are in the file group that has a USE "master". Here, it would correspond to a standardized representation of the Lotus document and the formerly encapsulated files. The Lotus document without its attachments, where all the descriptive and provenance information would remain, would, in its transformed version, make up a file group with the USE "documentation", which designates in SPAR the set of files containing metadata that cannot be entirely incorporated to the METS manifest but should be accessed for preservation planning. This document in its proprietary DXL format would be part of a new type of file group in SPAR, with the USE attribute "original": working with the designated community of the Administrative Channel has made the SPAR team realize that it lacked a USE in its nomenclature for files that are not the primary object of preservation but must be stored for reuse in their original environment.

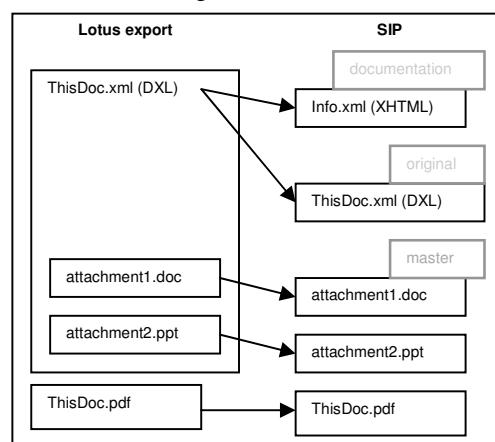


Figure 1. Creating a SIP from a Lotus Notes document

Using a similar logic, it appeared that in order to maintain usability of the Lotus documents in their original environment and to secure a higher service in the preservation process, attached files in proprietary formats could be transformed as well. This would be better accomplished not at the SIP creation stage, which deals with the way the Lotus export is recomposed, but within the system, according to the preservation planning capacities of SPAR at the time of the ingest. For example, a Microsoft Word binary file could be transformed into an Open Document file. The original Word file would be preserved in the information package for dissemination via the Lotus Notes interface, but would be

moved to the file group with the USE "original", while the new Open Document file would now be part of the file group with the USE "master", as the option chosen for long-term preservation actions (see Figure 2).

As for the DIPs, depending on the time and context of dissemination, they could combine files from file groups of different uses. This is yet another function that the SPAR team has had to take into account rapidly as a result of the dialogue with the representatives of producers and users in the Administrative Channel, since the repository so far can only disseminate DIPs that are an exact copy of the AIPs.

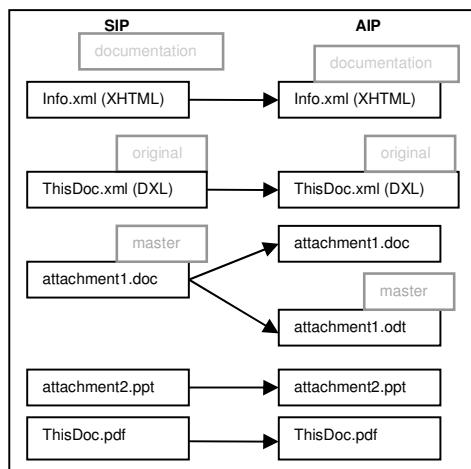


Figure 2. Migrating and moving files from SIP to AIP

3.2.3 Ending the Lifecycle: How to Delete

More flexibility at the access stage was something planned at the design stages of SPAR, that was scaled back because the communities for the first channels had no use for it, and moved forward again when producers and users made the case for its importance in their collection curation processes. Another example of these shifting priorities to serve the community is the deletion function. In the beginnings of the SPAR project, a lifecycle was devised for every AIP in the system: their first version, or version 0, would be preserved forever, as well as the latest one, and the one before, to allow for rollback. The implementation of this model was delayed, all the more since the first channels in SPAR contained collections whose forms were stable and was preservation was infinite.

Working with the records managers and their IT counterparts has shown the SPAR team that the deletion mechanisms have to be much more supple, while remaining simple, because of the high degree of human expert intervention in the lifecycle decisions. Although the documents in Lotus contain information regarding the duration of preservation required that is automatically assigned according to the document type, it cannot be used to pilot lifecycle decisions in SPAR: the intervention of an archivist to decide which documents are part of a closed case and are ready to be archived in the repository is necessary. Similarly, the BnF's archivists must validate all deletions. Moreover, these deletions have to be properly documented.

Given the design of SPAR, a solution might be to submit SIPs describing a "deletion request" event in their METS manifests. This would update the AIPs to include a "deletion processed" event documenting the action in their manifests while ridding them of their data objects, and set off the deletion of all previous versions of the AIPs. In any case, integrating such new and crucial

abilities into a functioning system will be an interesting challenge for the end of the year.

4. CONCLUSION: CURRENT ACHIEVEMENTS AND NEXT STEPS

4.1 Coverage of the OAIS Model

In its original conception, SPAR was intended to implement, as strictly as possible, of the OAIS model – indeed both OAIS models, the information and the functional models. Considering what has been achieved, to what extent has this objective been reached?

4.1.1 Information Model

The repository uses the full typology of information in the OAIS information model – but its precise nature, the way it is organized and the level at which it can be found highly differs from one track to another. In the digitization and audiovisual tracks, most metadata are recorded in the METS manifests. These METS files directly express structural metadata, and thanks to other metadata schemes embedded in METS, contain representation information (in MIX for images, textMD for text and MPEG-7 for audiovisual content), provenance and context information (in PREMIS), and descriptive information (mainly in Dublin Core). Fixity (checksums) and reference information (ISBN for books, persistent identifiers for all kind of documents, etc.) are included as well.

On the contrary, in the web legal deposit track, some representation information (MIME types of each contained file) is directly available in the ARC files, but is not described in METS. Moreover, METS files contain very few structural metadata, as the structure of web archives is already recorded in the hyperlinks present in the archived web pages. Descriptive information is only available at a very high level. In the end, it is perhaps in the use of PREMIS for context and provenance that the different tracks are the most similar.

As for rights metadata, which were not identified as such in the first version of the OAIS, they are not described yet in the metadata profiles. However, any descriptive, context or provenance information may be the basis for rights metadata, as they may help deduce the legal statuses of the documents. In fact, the very definition of each track depends on the legal status of the documents in it.

4.1.2 Functional Model

As to the functional model, one might consider that all functional entities have been implemented in SPAR modules – but at very different levels of completion. Modules highly related to collection knowledge and collection storage reached a high level of achievement: the ingest module extracts and computes a large number of metadata, which can be requested by the data management module. The storage and "storage abstraction services" modules are able to choose dynamically between different media storage and on what physical sites data should be stored. On the other hand, the access entity functional scope has been reduced to the bare minimum: to extract requested AIPs as they are from the system.

Yet the SPAR system has never been thought as a dark archive or a black box, but as an accessible system. However, designing a generic access module, able to create custom DIPs for digitized books, video games as well as web archives, is an objective currently beyond reach – and too ambitious for a project which was intended to show concrete results in a few years.

Finally, there is still work to be done on the administration and the preservation planning sides. New administration features are added each time new tracks and channels are developed, but a lot of improvements can be made on interfaces and ergonomics. These enhancements will probably be accelerated by the growing number of users as new challenges appear.

The preservation planning aspect is also less developed than what is expected in the OAIS model. On one hand, many functionalities of SPAR help design preservation strategies. Knowledge gathered at ingest, especially during identification and characterization processes, represents the cornerstone of a preservation strategy. On the other hand, we still do not have any tool to match automatically formats to preservation strategies. One of the next steps would be to let the system interact with format repositories like UDFR.

4.2 Next Steps

The second main phase of development will therefore extend the scope of SPAR in several directions:

- ingesting new types of collections. The administrative archives track is the next one to be integrated; electronic periodicals acquired by the BnF, e-books and other digital-born documents collected through legal deposit will have to follow.

- improving existing tracks, by adding new channels for instance. These new channels could be based, not only on the legal and technical statuses of the documents, but also on their scientific, heritage or financial value – taking into account the fact that this value may evolve through times.

- opening the repository storage and preservation facilities to the BnF's national partners using SPAR's third-party archiving track – in the heritage realm or not. This is probably less a technical than an organizational issue: to whom should these services be offered? At what cost? Who will be liable in case of problems?

- defining the professional profiles involved in the development and the daily use of SPAR. Until now, the development of the SPAR project has been followed on a day-to-day basis by two kind of professional profiles: IT engineers (developers and analysts) and “digital preservation experts”, i.e. librarians with a strong technical knowledge, who are in charge of assessing and maintaining metadata and data formats. Representatives of the Producers and User communities are also involved in the design stages of their tracks. However, a larger permanent working team is needed to maintain the live system while the developments continue. The content curators need to be more involved in the preservation of the collections they helped creating. Otherwise, digital collection curation and preservation will never be considered mainstream librarian activities.

The human part of digital preservation has probably been the least studied up to now, even though a working group called ORHION (Organization and Human Resources under Digital Influence) has been since 2009 dedicated to these issues [1 and 3]. A whole librarianship activity needs to be built around the SPAR system. Who will manage the system? Who will be able to send requests to the data management module? Who will be able to update metadata? Who will decide on preservation actions? This points to a general problem about the Designated communities and the frontier in their daily work between preservation and curation activities: is SPAR designed to be a digital curation tool as well as a preservation repository, or must new tools be developed as new needs are identified?

In its first design, SPAR was supposed to be a fully integrated digital preservation system. It is now a secure storage repository that offers its communities the ability to know and to manage all their digital collections. Some preservation actions happen outside SPAR – but the system is able to document them. On the other hand, SPAR makes a lot of information available for the first time, giving insight and control on the digital collections it holds. From this point of view, SPAR is redesigning the frontiers between preservation systems and curation tools at the BnF, reinventing librarianship for digitized and digital-born collections.

5. REFERENCES

- [1] Bermès, E. and Fauduet, L. 2010. The Human Face of Digital Preservation: Organizational and Staff Challenges, and Initiatives at the Bibliothèque nationale de France. *International Journal of Digital Curation*, 6, 1 (2011), 226-237.
[<http://www.ijdc.net/index.php/ijdc/article/view/175/244>]
- [2] Bermès, E. and Illien, G. 2009. Metrics and strategies for web heritage management and preservation. In *Proceedings of the 75th Congress of the International Federation of Library Associations* (Milan, Italy, August 23-27, 2009).
[<http://www.ifla.org/files/hq/papers/ifla75/92-bermes-en.pdf>]
- [3] Clatin, M., Fauduet, L. and Oury, C. 2012. Watching the library change, making the library change? An observatory of digital influence on organizations and skills at the Bibliothèque nationale de France. To be published in *Proceedings of the 78th Congress of the International Federation of Library Associations* (Mikkeli, Finland, August 11-17, 2012).
- [4] Cochrane, E. 2012. *Rendering Matters - Report on the results of research into digital object rendering*. Technical Report. Archives New Zealand. [<http://archives.govt.nz/rendering-matters-report-results-research-digital-object-rendering>]
- [5] Fauduet, L. and Peyrard, S. 2010. A data-first preservation strategy: data management in SPAR. In *Proceedings of the 7th International Conference on Preservation of Digital Objects* (Vienna, Austria, September 19-24, 2010).
[http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/fauduet_13.pdf]
- [6] Illien, G. and Stirling, P. 2011. The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future. In *Proceedings of the 77th Congress of the International Federation of Library Associations* (San Juan, Puerto Rico, August 13-18, 2011).
[<http://conference.ifla.org/past/ifla77/193-stirling-en.pdf>]
- [7] Oury, C. 2010. Large-scale collections under the magnifying glass: format identification for web. In *Proceedings of the 7th International Conference on Preservation of Digital Objects* (Vienna, Austria, September 19-24, 2010).
[http://netpreserve.org/about/Poster_ipres2010_webarchivefileformats_oury.pdf]
- [8] Oury, C. and Peyrard, S. 2011. From the World Wide Web to digital library stacks: preserving the French web archives, In *Proceedings of the 8th International Conference on Preservation of Digital Objects* (Singapore, November 1-4, 2011), 231-241.
[http://getfile3.posterous.com/getfile/files.posterous.com/temp_2012-01-02/dHqmzjcCGoexvmlBzJDCyhrlhIgswoffzvsfnPFAxjHFFEsarvwahEHrmyvj/iPRES2011.proceedings.pdf]

Developing a Community Capability Model Framework for data-intensive research

Liz Lyon

UKOLN, University of Bath

Bath BA2 7AY

United Kingdom

+44 1225 386580

e.j.lyon@ukoln.ac.uk

Alexander Ball

UKOLN, University of Bath

Bath BA2 7AY

United Kingdom

+44 1225 386580

a.ball@ukoln.ac.uk

Monica Duke

UKOLN, University of Bath

Bath BA2 7AY

United Kingdom

+44 1225 386580

m.duke@ukoln.ac.uk

Michael Day

UKOLN, University of Bath

Bath BA2 7AY

United Kingdom

+44 1225 383923

m.day@ukoln.ac.uk

ABSTRACT

Researchers across a range of fields have been inspired by the possibilities of data-intensive research. In many cases, however, researchers find themselves unable to take part due to a lack of facilities, insufficient access to data, cultural disincentives, and a range of other impediments. In order to develop a deeper understanding of this, UKOLN, University of Bath and Microsoft Research have been collaborating on developing a Community Capability Model Framework (CCMF) designed to assist institutions, research funding-bodies and researchers to enhance the capability of their communities to perform data-intensive research. This paper explores the rationale for using capability modelling for informing the development of data-intensive research and outlines the main capability factors underlying the current version of the CCMF.

Categories and Subject Descriptors

H.2.8 [Database Management]: Database Applications - Scientific databases

General Terms

Management, Measurement, Performance, Design, Economics, Human Factors

Keywords

Data-intensive research, Fourth Paradigm, capability modeling, research data, managing research data

1. INTRODUCTION

Following the publication of *The Fourth Paradigm* [1], researchers across a range of fields have been inspired by the

possibilities of data-intensive research, that is, research involving large amounts of data, often combined from many sources across multiple disciplines, and requiring some degree of computational analysis. In many cases, however, researchers find themselves unable to take part due to a lack of facilities, insufficient access to data, cultural disincentives, and a range of other impediments. In order to develop a deeper understanding of this, UKOLN, University of Bath and Microsoft Research have been collaborating on developing a Community Capability Model Framework (CCMF) designed to assist institutions, research funding-bodies and researchers to enhance the capability of their communities to perform data-intensive research by:

- profiling the current readiness or capability of the community;
- indicating priority areas for change and investment, and;
- developing roadmaps for achieving a target state of readiness.

In this paper, we will introduce the current version of the CCMF, outline some of the concepts underlying it and explain how it came to be in its current form.

2. DEFINITIONS

Data-intensive research belongs to what Gray [2] has termed the Fourth Paradigm of science, that is one primarily based on large-scale 'data exploration'. It is typified by workflows where researchers only apply their academic insight to data after an intense period of data collection and processing, with the processing stages dominant. Most 'big-science' disciplines - e.g., high energy physics, astronomy - are inherently data-intensive, while fields like the life sciences and chemistry have been utterly transformed in recent decades by the sheer quantity of data potentially becoming available for analysis [3]. Even the humanities and social sciences are not exempt from this 'data deluge,' e.g. with the emerging interdisciplinary fields of computational social science [4] and 'culturomics' [5].

One of Gray's key insights was that current data infrastructures were largely insufficient to deal with the vast amounts of data being produced [6, 7]. For example, Kolker, *et al.* [8, p. 142] comment that in the life sciences, "existing data

storage resources and tools for analysis and visualization lack integration and can be difficult to disseminate and maintain because the resources (both people and cyberinfrastructure) are not organized to handle them."

The CCMF is intended to provide a framework for analysing the capacity of communities - through institutions, research funding-bodies and researchers - to deal with data-intensive research. For the purposes of the CCMF, the following characteristics are necessary indicators of data-intensive research:

a) The research typically involves intense computational analysis of data.

b) The research typically involves analysis of large quantities of data, that is, more data than a research team could reasonably be expected to review without software assistance.

Also, if research involves combining data from several different sources, where the different source datasets have been collected according to different principles, methods and models, and for a primary purpose other than the current one, then it is likely to be classed as data-intensive research.

In terms of the CCMF, a *community* is broadly understood to be a set of people who share a particular location within the structure of an institution or society in general. Communities typically engage in both common and collective activities, and develop shared values, vocabularies, strategies and tactics [9]. In the particular case of academia, the term 'community' can apply at several different granularities: from the set of all academics and researchers, to disciplines such as physics or chemistry, or to narrow sub-disciplines such as organic crystallography [10, section 2.4.1]. It can also apply to the academics and researchers within a particular institution or department, or those working on a common project. In the context of the CCMF, the communities we are most interested in modelling are those defined by a discipline, a sub-discipline, or an institution.

3. CAPABILITY MODELS

Capability models are widely used by industry to help identify key business competencies and activities, helping to determine whether, how easily, and how well a given organization or community would be able, in theory and in practice, to accomplish a given task. The project team looked at a range of existing capability models in order to inform the development of CCMF, amongst them the Capability Maturity Model for Software and the Cornell Maturity Model for digital preservation, both of which have been used to explore data management requirements.

3.1 Capability Maturity Model for Software

A particularly influential capability model has been the Capability Maturity Model for Software (CMM) developed by the Software Engineering Institute at Carnegie Mellon University. This is concerned with evaluating the capability of an organisation to develop software on specification, on time and on budget [11]. CMM is a tool that can be used to appraise the current state of an organisation's processes, set targets for how it should be operating, and draw up a roadmap of how to achieve those targets. CMM defines five levels of software process maturity:

1. Initial - software process *ad hoc*, occasionally chaotic
2. Repeatable - basic project management processes established, some process discipline

3. Defined - software process for management and engineering is documented, standardized and integrated
4. Managed - detailed measures of process and quality are collected, software processes understood and controlled
5. Optimizing - incorporating continuous process improvement and innovation

More recently, CMM has been applied to research data management in two independent initiatives. For example, the Australian National Data Service (ANDS) [12] provides descriptions of the five levels of maturity for four key process areas: Institutional policies and procedures; IT Infrastructure; Support Services; Managing Metadata. The ANDS version of the model is much simpler than CMM itself, with narrative descriptions of maturity levels within each process area replacing the sets of key practices and common features. The focus is on higher education institutions, with the four process areas mapping neatly onto groups and services such as senior management, IT support, researcher support or staff development, and the library. The model freely acknowledges that not all organisations will aim to attain Level 5 (optimized) in all areas.

Crowston and Qin [13] take a different approach, focusing on scientific data management within research projects. They interpret the five levels as follows.

1. Data are managed within the project on an *ad hoc* basis, following the intuitions of the project staff.
2. Plans, policies and procedures are in place for data management, but they are peculiar to the project and reactive in nature.
3. The project tailors for itself plans, policies and procedures set up for data management at the discipline, community or institutional level; these plans tend to be pro-active in nature.
4. The project measures the success and effectiveness of its data management to ensure standards are maintained.
5. The project identifies weaknesses in its data management and addresses the defects proactively.

In developing their version of the model, Crowston and Qin consulted data management literature to identify key practices in data management, which they grouped into the following four key process areas:

1. Data acquisition, processing and quality assurance (3 practices)
2. Data description and representation (7 practices, including 'Develop and apply metadata specifications and schemas', 'Design mechanisms to link datasets with publications', 'Ensure interoperability with data and metadata standards')
3. Data dissemination (4 practices, including 'Encourage sharing', 'Distribute data')
4. Repository services/preservation (7 practices, including 'Store, backup and secure data', 'Perform data migration', 'Validate data archives')

In addition, they identified several generic practices that closely resembled those in the earlier models, for example:

developing policies for data release, sharing, data rights and restrictions, and data curation; identifying staffing needs; developing business models; developing data management tools; training researchers and support staff; capturing provenance data; developing collaborations and partnerships; assessing impact and enforcing policy.

The use cases for all of these capability models strongly resemble those intended for the CCMF. They provide a clear framework for characterising an organisation or project, and identifying improvements that could be made as well as the order in which they should be tackled. They also provide a reference vocabulary for describing relevant activities and functions, without being overly specific about how these should be carried out or implemented. While CMM is primarily focused on the commercial sector, the version of the model developed by ANDS shows, however, how it can be applied to higher education institutions. Crowston and Qin's model focuses on research projects while also referencing (and having clear implications for) the wider institutional and disciplinary context. Indeed, perhaps the most important difference to reconcile between these models and what is required for the CCMF is that they again admit only one target state to which organisations should aspire, with the possible exception of the ANDS model; in contrast, it would be difficult to find a single generic description that could apply to all successful forms of data-intensive research.

3.2 Cornell Maturity Model

A slightly different approach to capability modelling was developed in the Cornell Maturity Model used to analyse the type of response given by higher education institutions to the challenges of digital preservation. Kenney and McGovern [14, 15] present a distinctive five-stage maturity model:

- Acknowledge. The institution recognises it must perform some degree of digital preservation.
- Act. The institution instigates digital preservation projects.
- Consolidate. The institution embeds digital preservation as ongoing programmes.
- Institutionalise. The institution unifies the various digital preservation activities into a single programme.
- Externalise. The institution collaborates with others to achieve economies of scale and increased digital preservation capability.

In the early expressions of the Cornell model, key indicators for each stage were described along the three dimensions of policy and planning, technological infrastructure, and content and use. These dimensions were later changed to organisational infrastructure, technological infrastructure, and resources, with a corresponding new set of key indicators. To emphasise that organisations should develop in each of the dimensions in parallel, but that the digital preservation capability can still be stable with uneven development, they became known as the three legs of a digital preservation Three-Legged Stool, with legs for organization, technology and resources.

The Cornell model was further developed by the JISC-funded AIDA Project into a scorecard-based tool for benchmarking the current state of digital asset management within institutions or departments. AIDA expanded and formalised the indicators within each leg, arriving at eleven metrics in each of

the organisation and technology legs, and nine metrics within the resources leg. While AIDA was intended as a self-assessment toolkit, the AIDA Project Team provided a service for assessing completed scorecards to determine an overall picture of institutional readiness, recommend actions for increasing readiness, and provide guidance on digital asset management issues.

The AIDA scorecard provided by the Project Team was in the form of a Microsoft Word document with form controls, with analysis performed on an accompanying Excel spreadsheet. The process of performing the benchmarking exercise itself, though, was left up to the individual to plan. Sensing a need, the UK Digital Curation Centre (DCC) applied its experience from developing the tools that supported DRAMBORA and the Digital Asset Framework (DAF) to produce a Web-based tool allowing a team of contributors to collaborate on an AIDA-style self-assessment. This tool, known as CARDIO [16], uses a very similar set of metrics ('statements') to those developed by AIDA, but has a specific emphasis on research data and can be used at multiple levels of organizational granularity (project, department, institution).

The use cases for this model – assessing the current state of readiness of an institution and identifying priorities for development – again resonate strongly with those for the CCMF. Just as the CCMF should be applicable to researchers, institutions and funding bodies, the Three-Legged Stool can be applied at several different granularities. The notion of having broad, abstract dimensions measured according to specific, concrete metrics is a useful one. Once more, though, the model considers only one correct route from nil readiness to complete readiness through each leg, and through each metric within each leg. The CCMF, by contrast, needs to model several types of community capability and - by implication - several different 'routes' to achieving capability.

4. CCMF CAPABILITY FACTORS

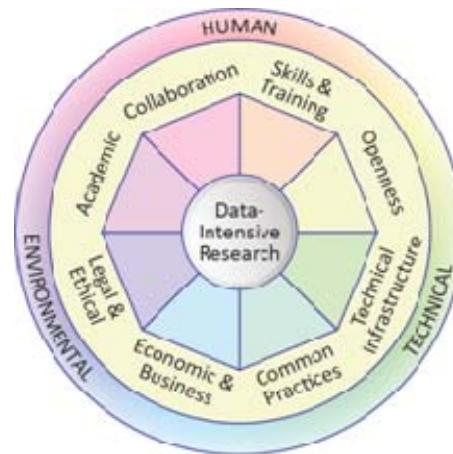


Figure 1: Community Capability Model Framework

We propose a Community Capability Model Framework for data-intensive research comprising eight capability factors representing human, technical and environmental issues (Figure 1). Within each factor are a series of community characteristics that we feel are relevant for determining the capability or readiness of that community to perform data-intensive research. In this section, we

will outline the eight capability factors that make-up the CCMF and comment on some of the characteristics associated with each one. The CCMF consultation draft [17] provides additional detail on all of these, including:

- an identification of the community characteristics associated with each factor, including indications of how each characteristic could be 'measured' for the purposes of analysis and comparison;
- one or more exemplars demonstrating how the alternatives should be interpreted, and;
- brief commentary explaining the relevance of the characteristic for determining capability, and how the project team's thinking has been shaped by the literature and by discussions with the community. These discussions took place in a series of five workshops held between September 2011 and February 2012 in the UK, US, Sweden and Australia.

4.1 Collaboration

The working relationships that are formed during research have a strong bearing on the types of research that can be performed. Collaborations can be informal or semi-formal, or can be rigorously controlled, managed and enforced through contracts and agreements. Collaboration can be organised within a discipline, between two or more disciplines, with organizations outside the research sector, and with the general public.

4.1.1 Collaboration within the discipline/sector

The level of collaboration within a discipline can range from almost none (sometimes characterised as the lone researcher) to extremely large, well-organised national or international consortia. In practice, however, perhaps most disciplinary collaboration is focused on a particular research group or groups. For example, bioinformatics and neuroinformatics are dominated by small teams, with relatively few large-scale contributors. By contrast, big science disciplines like high energy physics and astronomy are typically organised in projects at international scale.

It is recognised that individual researchers can move along the spectrum as their career progresses, e.g. first working alone on an idea or hypothesis, exposing it gradually to colleagues and gaining collaborators from the research group and, at a later stage, the wider community.

4.1.2 Collaboration/ interaction across disciplines

Interdisciplinary collaborations follow the same broad pattern as those within disciplines. Some disciplines will have next to no interaction with others while others will have forged formal collaborations over relatively long periods of time.

Interdisciplinarity is one response to the perceived overspecialisation of research disciplines, and can be encouraged in institutional or national contexts through the creation of matrix structures like joint research centres or faculty appointments [18, pp. 173-4]. Data-intensive research will tend towards the interdisciplinary, not least because it requires the input of computational specialists. There are many potential impediments to interdisciplinary collaboration, not least epistemic barriers based upon what Jacobs and Frickel [19, p. 47] describe as "incompatible styles of thought, research traditions, techniques, and language that are difficult to translate across disciplinary domains."

4.1.3 Collaboration/ interaction across sectors

Researchers will sometimes need to collaborate across sector boundaries, e.g. with industry, equipment suppliers, media, professional bodies or public sector organisations. The types of organization suitable for collaboration will vary quite widely, and might include: pharmaceutical and biotechnology companies (in medicine and the life sciences), natural history museums (in biodiversity, ecology and palaeontology), or the digital content industries (e.g., Google Book Search for culturonomics).

4.1.4 Collaboration with the public

There is a growing interest in public engagement with research. This is particularly strong in the life sciences, where some funding bodies (e.g., medical research charities) are keen to involve patients in things like reviewing grant proposals. In fields as divergent as astronomy (GalaxyZoo) and papyrology (Ancient Lives), members of the public are being encouraged to contribute directly to some aspects of the research process.

4.2 Skills and training

The capability of a community to perform data-intensive research is strongly influenced by the individual capabilities of its members, and the capacity that results from the combination and multiplication of these capabilities. Community capability can therefore be enhanced by training members in the relevant skills. This training is most effective when it is fully embedded as part of the early education and continuing professional development of researchers.

4.2.1 Skill sets

The capability of a community to perform data-intensive research is strongly influenced by the individual capabilities of its members, and the capacity that results from the combination and multiplication of these capabilities. Community capability can therefore be enhanced by training members in the relevant skills. This training is most effective when it is fully embedded as part of the early education and continuing professional development of researchers.

4.2.2 Pervasiveness of training

There is much variation across disciplines, institutions and degree programmes in the provision of training. Some UK research funding bodies have established Doctoral Training Centres to develop and deliver training programmes for their disciplinary communities. JISC has funded training materials that target particular disciplines e.g. psychology. At some institutions - including the University of Bath - support services like subject liaison librarians and IT services are beginning to develop a range of training programmes for researchers, covering topics such as data management planning. The UK Digital Curation Centre has delivered training modules on a regional basis as part of its Regional Roadshow Programme, while national data centres such as the ESDS (in the UK) and ICPSR (in the US) run workshops on data management.

4.3 Openness

Historically, scientific progress has been driven forward by the open communication of research methods and results. More generally, the principle of openness can be applied at different levels: from openness in communicating the plans for research and ongoing progress whilst the research is undertaken, to opening up the published literature to a wider audience. Driven by concerns for improving the validation, reproducibility and

reusability of research, the last decade has also seen calls for opening up the data and other details of methodologies employed, alongside final results and conclusions, for scrutiny and re-use by the wider community, a process that is considered by some to add value to research.

4.3.1 Openness in the course of research

This characteristic describes whether researchers choose to communicate information about their research whilst it is still ongoing, the extent to which they make their plans and intermediate results known, and the mechanisms they use to achieve openness. Such openness makes an informal variety of early peer review possible, which in the long term may result in more interoperable data and therefore more opportunities for data-intensive research.

4.3.2 Openness of published literature

The body of published literature can be available under different conditions – some literature is available only through payment agreements; sometimes only the description of the literature (metadata) is accessible, whilst at the other extreme some communities have embraced the practice of sharing of all the published literature through archives freely available to all. The openness or otherwise of a publication may depend on its type (journal paper, conference paper, thesis), or readers may need to make specific personal requests in order to gain access. Providing open access to published literature may make it easier for potential re-users to locate suitable data.

4.3.3 Openness of data

There is wide variation in the openness of data. In some disciplines, e.g. astronomy, proteomics and philology, data is routinely published openly, sometimes after a period of exclusive use. In others, there is no tradition of data sharing. For example, O'Donoghue, *et al.* [20] note the unevenness of availability of biological data, with the two extremes exemplified by PDB, which contains almost all experimentally determined structures, and image data from high throughput experiments, where there is little data integration and 'most of these data are never made publicly available'.

Treloar [21] presents a model of data openness with the following three categories:

1. Private research domain. Typically access is tightly controlled and restricted to a core team within a single institution. Technological platforms such as laboratory information management systems or research management systems are used.
2. Shared research domain. This is where some, but not all, the data is shared by the core team with other colleagues, often outside the home institution.
3. Public domain. Data is published so that (with a few exceptions) anyone can gain access to it. Institutional repositories may be used to provide this access. Typically the data will be given a persistent identifier, and the associated metadata will be fixed.

4.3.4 Openness of methodologies/workflows

Releasing data alone may not be sufficient to replicate results and findings. Details of methodologies and workflows which allow other researchers to reproduce the workings and methods of other groups may be required. This characteristic describes the practice

of sharing information regarding the processes employed, either as descriptions or in executable forms, so that one researcher can apply the same methods either to the same dataset or perhaps to alternative data or applications.

4.3.5 Reuse of existing data

This characteristic focuses on the attitudes and practices of using data sets generated by other researchers. Researchers may be open to regularly using data shared by others, but they may only trust specific sources. Data sets obtained from the community can be processed in different ways – data can be aggregated, re-analysed under the original conditions or mined to generate new insights.

4.4 Technical infrastructure

The technical infrastructure that supports research comprises tools and services that are used at different stages of the research life cycle. This capability factor describes categories of tools and services that meet user needs across various activities.

4.4.1 Computational tools and algorithms

Computational tools and algorithms form the backbone of most data-intensive research workflows. If such tools under perform, it places a hard limit on what research can be conducted.

4.4.2 Tool support for data capture and processing

Tools that support data capture and processing often make assumptions about the formats in which the data is stored and processed. The extent to which the tools support formats that are more widely supported by other tools may determine whether data can be shared, understood, processed and re-used within the wider technical environment. When the tools support open or agreed formats or the interchange of data in different formats, tool interoperability increases.

4.4.3 Data storage

Data storage needs to grow as data volumes increase, but requirements may also be defined by the data type. Such requirements may involve issues of physical location, performance, access control and security, scalability, reliability, and speed as well as capacity. For example, in some communities the storage of clinical data must adhere to the ISO/IEC 27000 series of information security standards. Data storage can be organised locally, nationally or globally. Interactions with data storage are required by several of the other tool categories, such as data capture and processing tools, discovery services and curation and preservation services.

4.4.4 Support for curation and preservation

The relative importance of the tools that enhance contemporary usefulness of data and those that aid its long-term preservation varies between disciplines. For disciplines reliant on non-replicable observations, good preservation tools help to maintain stocks of data for future data-intensive research.

4.4.5 Data discovery and access

Data discovery and access is currently problematic because different types of catalogues do not integrate well and there is no standard way to publish them, and no easy way to federate them for cross-discovery. Other challenges exist at the semantic level [22, 23]. One measure suggested would be to see how far a community might be from agreeing standards.

4.4.6 Integration and collaboration platforms

Integration and collaboration tools may help researchers manage their workflows and interactions more efficiently, increasing their capacity for data-intensive research.

4.4.7 Visualisations and representations

Visualisation tools are extremely important for data-intensive science. However, the current range of visualisation tools tends to be fragmented and not necessarily optimized for the scales of data becoming available [20].

4.4.8 Platforms for citizen science

Citizen science platforms provide infrastructure that enables non-specialists to participate and collaborate in the research process. Whilst the platforms can be developed within a specific project they can then be redeployed to meet the need of other communities.

4.5 Common practices

This capability factor describes community practices that have produced standards, whether by design or *de facto*. The quantity of standards in a particular discipline is not necessarily a measure of its capability. In some cases, standards may actually hold back progress, especially where they are poorly supported by software or where competing standards effectively act as data silos. It is the quality of data standards that is important, specifically whether they promote and enable the re-use and combination of data. While convergence on a *de facto* standard can happen organically, designed standards typically need to be driven either by influential organisations at a national or international level, or else by a dedicated and enthusiastic association of individuals within a community.

4.5.1 Data formats

These are formats that describe how data is encoded and stored, and facilitate data exchange.

4.5.2 Data collection methods

Data collection methods can also be standardised and shared. Methods are varied depending on the activity within which collection is undertaken. Data collection activities include observational collection, instrumental collection requiring calibration, survey data, sensor data and performance data.

4.5.3 Processing workflows

If data has been processed according to standard and accepted workflows, it is more likely to be considered for reuse by other researchers.

4.5.4 Data packaging and transfer protocols

Agreed standards for data packaging and transfer ease the transport of data between creators, archives and the re-users of data.

4.5.5 Data description

Data description standards are used to make data re-usable by providing metadata that describes different aspects of the data. Whilst some disciplines have adopted description schemes that become widely used, other schemes are at earlier stages of adoption and have not yet fulfilled the promise of data interoperation and reusability that they are intended to facilitate. Schemes can be aimed at a generic level or be specialised with discipline-specific fields.

4.5.6 Vocabularies, semantics, ontologies

Vocabularies, semantics and ontologies are also used by communities to exchange information and data, and attempt to capture the knowledge, concepts and terminologies within the discipline in a standardised agreed format. Some are adopted within specialised communities, whilst others find their place as a bridge between communities. Different models for how these standards are agreed and maintained can be described, and their progression or maturity follows a trajectory from proposal and specification to standardisation by recognised bodies.

4.5.7 Data identifiers

Data identifiers are developed to provide unique and unambiguous methods to refer to or access research objects. They may serve the purposes of identification and location. The objects may be literature, chemical or biological entities, or entries in databases.

4.5.8 Stable, documented APIs

Where data repositories and data processing services provide APIs, it opens up the possibilities for automated workflows and thereby increases the scale at which research can be performed.

4.6 Economic and business models

Moving into data-intensive research requires some degree of investment, and it is therefore important to consider how this might be funded and the business case for making the move. Disciplinary differences are important here: the business case will be easier to make where it is important to publish quickly and generate many research papers from a single investment, and harder where the emphasis is on careful and considered weighing of evidence.

4.6.1 Funding models for research and infrastructure

There are many thematic perspectives to consider here including scholarly communication and data publishing models, approaches to data curation and preservation, network-level infrastructure, through to capacity-building programmes. The established political and funding landscape in a particular geographical area is strongly influential in determining the business models in place. In order to realise the full potential global scale of data-intensive research, politico-legal issues and barriers linked to trans-national borders, will need to be overcome.

4.6.2 Public–private partnerships

In communities where it is common for research to be partially or wholly funded by the private sector, the diversity of funding streams may make the research more sustainable, and the research may have greater impact outside academia. At the same time, the research may be contingent on business models and return on investment, and it is less likely that data will be made available for reuse.

4.7 Legal and ethical issues

Quite apart from any cultural barriers that may obstruct data sharing, and thereby restrict the scope for data-intensive research, in some cases there may be ethical reasons why certain datasets may not be shared, and legal barriers both to sharing data in the first place and to recombining it for the purposes of data-intensive research. Even in cases where the barriers do not in fact exist, ambiguities and misperceptions of the legal or ethical position may deter risk-averse institutions and researchers from pursuing

such lines of enquiry. It will, therefore, be easier for data-intensive research to flourish where the legal issues surrounding data sharing and reuse are well understood and well managed, and where there are established frameworks for ensuring such research is conducted in an ethical manner.

The following characteristics should be assessed with caution, as the official policies do not always reflect what is actually done by researchers and institutions.

4.7.1 Legal and regulatory frameworks

At issue here are laws that impact on the sharing and reuse of data (most notably intellectual property laws and contract law), as well as relevant policies and regulations adopted by governments, funding bodies, professional societies and other bodies. The benefit of legal and regulatory frameworks for community capability lies in the clarity they provide with respect to the law, so that it is readily apparent whether and how data may be shared and reused. In the UK, such frameworks might, for example, instruct researchers to record the owner of data, to avoid future uncertainty over the contractual arrangements under which the researcher was working. There are several points of failure, though, that must be avoided. No framework will be able to work around firm legal prohibitions. In some US jurisdictions there are limitations on state-based contracts, signing contracts outside of the state, and selling outside the state by state-based institutions. Where the law itself is ambiguous or untested, any framework for managing compliance will necessarily be cautious. More helpful frameworks may build on the firmer parts of the law to allow routes for data sharing and reuse, while more obstructive frameworks might block the possibility entirely. Even where helpful frameworks do exist, researchers must be familiar with them and trust them. Funding bodies, professional societies, governing bodies and regulators play a large part in ensuring adherence to procedures and community norms, but their attitudes may not always be favourable to the needs of data-intensive research.

4.7.2 Management of ethical responsibilities and norms

As with the previous characteristic, the issue here is with clarity. Researchers will feel more confident about releasing sensitive data if there are established and trusted procedures in place for anonymising it, limiting access to it, and so on. There are also ethical issues relating to research quality.

4.8 Academic culture

The community norms that exist for the process of doing research are a key factor in determining the level of support a researcher might expect when moving into data-intensive research. Such a move may be easier where entrepreneurship and innovation are welcomed, and harder where such things are frowned upon. Even more importantly, data-intensive research is most likely to flourish in communities where data is valued highly, where researchers are rewarded for their data contributions, and where high standards are expected of data entering the research record.

4.8.1 Productivity and return on investment

The impact that this characteristic has on community capability is relatively weak but it is still important to recognise. While the metric is couched in terms of timescales and publishing patterns, the underlying feature we are interested in is the character of the research. The rapid-cycle end of the dimension is the natural home for disciplines where the interest is in finding new things:

new particles, new molecules, new sequences. The slow-cycle end of the dimension is the natural home for disciplines where the interest is in profound insight, and improved understanding of complex issues. Data-intensive research methods can assist in all these areas of enquiry, but the immediacy of their impact varies. At the rapid-cycle end, it is relatively straightforward to decide which patterns to look for in data, and relatively obvious when an interesting result has been found; in such cases an investment in the means of data-intensive research has a quick pay-off. At the slow-cycle end, it is typically harder to assemble a comprehensive dataset to analyse, and the analytical steps to automate may themselves require debate and justification; in such cases, greater preparation is needed before data-intensive research methods are applied, and once they are it may take some time to reap the benefits.

4.8.2 Entrepreneurship, innovation and risk

The move to a new paradigm of research requires a certain degree of investment, in both time and effort, and there is always a risk that it may not produce interesting results, or that peer reviewers may not accept the new methodology. There is therefore risk to both PIs and funding bodies when it comes to funding such research. In disciplines where risk-taking and innovation are seen in a positive light, this is less of a barrier.

4.8.3 Reward models for researchers

Contributions to data intensive research are made in different ways. Not all these contributions are formally recognised when considering rewards for researchers. Rewards can come in many forms including career advancement, recognition by peers and funding, both for research and for training students and junior researchers. Methods for measuring contributions are also varied, with some measures for example publications being well established. Even within publications, however, there are different ways of recording contribution. Multi-author efforts can credit each contributor. Other categories of contribution encompass software products and sharing of analysed data, such as DNA sequences. Some contributions such as efforts to curate data and make it reusable are notorious for being poorly recognised and rewarded.

4.8.4 Quality and validation frameworks

Even if data is shared, it may not be in a state amenable to reuse, let alone full validation. Unless data is sufficient quality, and provably so, it is of limited use in data-intensive research conducted by other researchers. A community's capability for such research, therefore, is increased where data is available that has been through thorough independent quality checks, and where this data is maintained and integrated with similar data by specialist curators.

5. CONCLUSIONS

The Community Capability Model Framework is a tool for evaluating a community's current readiness to perform data-intensive research, and for identifying areas where changes need to be made to increase capability. This paper has outlined the eight capability factors identified, which deal with human, technical and environmental issues. The detailed CCMF [17] attempts to identify characteristics that can be used to judge community capability.

While the CCMF has been developed with the involvement of a wide range of stakeholders and interested parties, the immediate next step will be to validate it by applying the

framework to a number of research communities. In the longer term we hope to develop tailored versions of the framework for different stakeholders, and to improve the usefulness of the tool as an aid to decision making and planning.

6. ACKNOWLEDGMENTS

The CCMF was developed as part of the Community Capability Model for Data-Intensive Research project, a partnership of Microsoft Research Connections and UKOLN, University of Bath: <http://communitymodel.sharepoint.com>

The authors would like to thank all those that attended CCMDIR workshops for their participation in the development process and for their comments on earlier drafts of the model.

7. REFERENCES

- [1] Hey, T., Tansley, S. and Tolle, K. Eds. 2009. *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research, Redmond, WA.
- [2] Gray, J. 2009. Jim Gray on eScience: a transformed scientific method. In *The fourth paradigm: data-intensive scientific discovery*, T. Hey, S. Tansley and K. Tolle, Eds. Microsoft Research, Redmond, WA, xix–xxxiii.
- [3] Hey, T., and Trefethen, A. 2003. The data deluge: an e-science perspective. In *Grid computing: making the global infrastructure a reality*, F. Berman, G. C. Fox, and T. Hey, Eds. Wiley, New York.
- [4] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., et al.. 2009. Computational Social Science. *Science* 323 (6 Feb), 721-723. DOI= <http://dx.doi.org/10.1126/science.1167742>
- [5] Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, T. G. B., et al. 2011. Quantitative analysis of culture using millions of digitised books. *Science* 331 (14 Jan), 176-182. DOI= <http://dx.doi.org/10.1126/science.1199644>
- [6] Gray, J., Liu, D. T., Nieto-Santisen, M., Szalay, A., DeWitt, D. J., and Heber, G. 2005. Scientific data management in the coming decade. *ACM SIGMOD Record* 34, 34-41. DOI= <http://dx.doi.org/10.1145/1107499.1107503>
- [7] Szalay, A., and Blakeley, J. A. 2009. Gray's Laws: database-centric computing in science. In *The fourth paradigm: data-intensive scientific discovery*, T. Hey, S. Tansley and K. Tolle, Eds. Microsoft Research, Redmond, WA, 5-11.
- [8] Kolker, E., Stewart, E., and Ozdemir, V. 2012. Opportunities and challenges for the life sciences community. *OMICS: A Journal of Integrative Biology* 16, 138-147. DOI= <http://dx.doi.org/10.1089/omi.2011.0152>
- [9] Agre, P. 1998. Designing genres for new media: social, economic, and political contexts, In *Cybersociety 2.0: revisiting computer-mediated community and technology*, S. G. Jones, Ed. SAGE, Thousand Oaks, CA, 69–99.
- [10] Treloar, A. 1998. Hypermedia online publishing: the transformation of the scholarly journal. PhD thesis, Monash University, Melbourne. <http://andrew.treloar.net/research/theses/phd/index.shtml>
- [11] Paulk, M. C., Curtis, B., Chrissis, M. B., and Weber, C. 1993. *Capability maturity model*, Version 1.1. Technical Report, CMU/SEI-93-TR-024 ESC-TR-93-177. Carnegie Mellon University, Software Engineering Institute, Pittsburgh PA. <http://www.sei.cmu.edu/reports/93tr024.pdf>
- [12] Australian National Data Service. 2011. *Research Data Management Framework: Capability Maturity Guide*. ANDS Guides. <http://ands.org.au/guides/dmframework/dmf-capability-maturity-guide.html>
- [13] Crowston, K. and Qin, J. 2012. A capability maturity model for scientific data management: evidence from the literature. *Proceedings of the American Society for Information Science and Technology* 48, 1-9. DOI= <http://dx.doi.org/10.1002/meet.2011.14504801036>
- [14] Kenney, A. R., and McGovern, N. Y. 2003. The five organisational stages of digital preservation. In *Digital libraries: a vision for the 21st century*, P. Hodges, M. Sandler, M. Bonn, and J. P. Wilkin, Eds. University of Michigan Scholarly Publishing Office, Ann Arbor, MI. <http://hdl.handle.net/2027/spo.bbv9812.0001.001>
- [15] Kenney, A. R., and McGovern, N. Y. 2005. The three-legged stool: institutional response to digital preservation. 2nd Convocatoria del Coloquio de marzo, Cuba, March. http://www.library.cornell.edu/iris/dpo/docs/Cuba-arknym_final.ppt
- [16] Digital Curation Centre, CARDIO: <http://cardio.dcc.ac.uk/>
- [17] Lyon, L., Ball, A., Duke, M., and Day, M. 2012. *Community Capability Model Framework* (consultation draft). UKOLN, University of Bath, Bath. <http://communitymodel.sharepoint.com/Documents/CCMDIRWhitePaper-v1-0.pdf>
- [18] National Academy of Sciences. 2004. *Facilitating interdisciplinary research*. National Academies Press, Washington, DC.
- [19] Jacobs, J. A., and Frickel, S. 2009. Interdisciplinarity: A Critical Assessment. *Annual Review of Sociology* 35 (2009), 43-65. DOI= <http://dx.doi.org/10.1146/annurev-soc-070308-115954>
- [20] O'Donoghue, S.. I., Gavin, A. -C., Gehlenborg, N., Goodsell, D. S., Hériché, J. K., North, C., et al. 2010. Visualizing biological data – now and in the future. *Nature Methods*, 7, S2–S4. DOI= <http://dx.doi.org/10.1038/nmeth.f.301>
- [21] Treloar, A. 2011. Private research, shared research, publication, and the boundary transitions. http://andrew.treloar.net/research/diagrams/data_curation_continuum.pdf
- [22] Bowker G. C. 2001, Biodiversity dataversity. *Social Studies of Science* 30, 643-84. DOI= <http://dx.doi.org/10.1177/030631200030005001>
- [23] Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C., and Borgman, C. L. 2011. Science friction: data, metadata, and collaboration. *Social Studies of Science* 41, 667-690. DOI= <http://dx.doi.org/10.1177/0306312711413314>

CRISP: Crowdsourcing Representation Information to Support Preservation

Maureen Pennock

The British Library
Wetherby
West Yorkshire
0044 1937 546302

maureen.pennock@bl.uk

Andrew N. Jackson

The British Library
Wetherby
West Yorkshire
0044 1937 546602

andrew.jackson@bl.uk

Paul Wheatley

University of Leeds
Leeds
West Yorkshire
0044 113 243 1751

p.r.wheatley@leeds.ac.uk

ABSTRACT

In this paper, we describe a new collaborative approach to the collection of representation information to ensure long term access to digital content. Representation information is essential for successful rendering of digital content in the future. Manual collection and maintenance of RI has so far proven to be highly resource intensive and is compounded by the massive scale of the challenge, especially for repositories with no format limitations. This solution combats these challenges by drawing upon the wisdom and knowledge of the crowd to identify online sources of representation information, which are then collected, classified, and managed using existing tools. We suggest that nominations can be harvested and preserved by participating established web archives, which themselves could obviously benefit from such extensive collections. This is a low cost, low resource approach to collecting essential representation information of widespread relevance.

Categories and Subject Descriptors

H.3.m [INFORMATION STORAGE AND RETRIEVAL]:
Miscellaneous

General Terms

Management, Documentation, Design, Experimentation, Human Factors, Verification.

Keywords

Representation information, crowdsourcing, digital preservation, web archiving, community engagement, social networking.

1. INTRODUCTION

Representation information (RI) is widely acknowledged as essential for digital resources to remain accessible into the future. The internet is one of the best sources of representation information, which is scattered around web in a variety of personal and organizational websites. Yet finding and navigating this information is not straightforward. We know from experience that the identification and collection of RI is highly resource

intensive. Organizations collating and maintaining resources themselves have struggled to resource this work. The PADI site remained a key source of information on digital preservation for a number of years but was eventually closed and web archived when the overhead of maintaining the information became too great. Furthermore, we know all too well that websites themselves are far from permanent. Vital online information about preservation tools and file formats can be transitory: here one day, 404'd the next.

Existing online community-created resources that link to online representation information sources go some way to addressing these challenges, though they are typically spread around quite thinly, with much duplication. A number of formal RI registries have been built but are sparsely populated, despite widespread community acceptance of the importance of RI, and there appears no overall consensus on the extent of RI required to support long term preservation and access.

The scale of this challenge requires a coordinated and collaborative effort across the wider preservation and curation communities, to establish an inclusive and (semi-)automated solution for RI collection and preservation. Encouraging more coordination will reduce duplication of resources and maximize effort in creating and maintaining the resources we need to make preservation effective.

2. DEFINING SHARED REPRESENTATION INFORMATION REQUIREMENTS

Representation information facilitates the proper rendering and understanding of content. In OAIS terms, RI is a distinct type of information object that may itself require representation information [1]. It can exist recursively until the knowledge base of the designated community dictates no further RI needs be recorded. As a result, the extent, size and boundaries of an RI collection are potentially immense. The vague boundaries and immense potential scope of an RI collection may be one of the reasons why RI collections have been so difficult to establish. We contend that the precise scoping of a core RI collection is the key to maximizing community input and establishing a successful well-populated collection. 'Core shared RI' is that which is most broadly relevant to the widest possible user base.

Brown, in his 2008 white paper on Representation Information Registries, defines two classes of structural RI: Descriptive and Instantiated [2]. These are defined respectively as information that describes how to interpret a data object (e.g. a format

specification) and information about a component of a technical environment that supports interpretation of the object (e.g. a tool or platform).

Descriptive structural RI such as format specifications, which are universally relevant for all objects of a given format regardless of the environment in which content has been used, are core shared RI. These are therefore our starting point for a core shared RI collection. We consider tools that support interpretation to be secondary shared RI, as whilst they are essential, their relevance is more likely to differ for different collecting institutions.

Format specifications are not just necessary for future access, but also contemporary preservation planning. The current SCAPE (Scalable Preservation Environments) project¹, funded by the EU, needs to collect format information to assist preservation planning and other processes. It is clear that the number of stakeholders with a vested interest in contributing to a shared format specification registry is extensive.

3. CURRENT INITIATIVES

The case for representation information has been well made elsewhere and will not be repeated here [3]. Numerous online RI resources have been established by the preservation community, each with slightly different foci, granularity and coverage. Here we introduce some of the key current resources.

3.1 Format registries

Several different format registry initiatives have been established in the preservation community over the past decade. These are now roughly consolidated into two initiatives: the UDFR and the proposed OPF format registry.

UDFR combines content previously collected in PRONOM and GDFR in a single, shared semantic registry [4]. Functional development is led by use cases. The system is highly structured with a well-defined ontology. It is publicly available and awareness of the resource is high, though the contributor base appears relatively low.

The proposed OPF format registry ecosystem will link existing sources of representation information and enable users to create linked data collections based on the information currently distributed across disparate resources [5]. Proposed components include the PLANETS core registry and PRONOM, in conjunction with a proposed ‘registry of registries’. The success of the project is dependent upon successful population of supporting registries.

Whilst both are labeled ‘registries’, a corresponding repository element is typically able to store RI directly.

3.2 Tool registries

A number of tool registries have been established and shared across the digital preservation community. The following list is not exhaustive but exemplifies the range and scope of currently available online tool resources.

The Digital Curation Centre (DCC) Tools & Services site identifies and links out to a large number of curatorial tools for deposit/ingest, archiving/preserving, and managing/administering repositories.² Many of the tools were developed by and are well established in the preservation community. The site is managed by

¹ SCAPE project website: <http://www.scape-project.eu/>

² DCC Tools & Services resource:
<http://www.dcc.ac.uk/resources/external/tools-services>

the DCC, though community nominations are encouraged by email.

A community wiki of precision digital preservation tools is provided by the OPF through the OPF Tool Registry.³ This includes tools developed in the AQuA and SPRUCE mashups, as well as the SCAPE project.⁴ Tools are categorized by function and simple user experiences described. Source code for some of the tools is hosted directly on the wiki. The site is manually populated by a small geographically distributed group of digital preservation professionals. Membership of the group is open to all, and all members have editing rights.

The Digital Curation Exchange Tool list is a flat though extensive list of links for tools and services relevant to digital preservation.⁵ It includes many ‘supporting’ services and developer tools absent from other lists, such as storage solutions, core utilities, and office plug-ins. Description is minimal. The list is maintained by the membership, which is open to all.

Finally, an inventory of Partner Tools & Services is available from the NDIIPP website, which briefly describes and shares information about tools and services used in NDIIPP.⁶ Entries are not categorized though the context of use is clearly identified. Some content is hosted directly on the site though many entries point to external links.

3.3 Other initiatives

The Library of Congress’ (LoC) Digital Formats Sustainability site contains extensive format descriptions relevant to the LoC collection.⁷ Format versions have their own entries. Descriptions link to format specifications published online and identify sustainability issues. Format specifications published on these pages are harvested by the LoC web archiving program. The site is maintained by LoC staff though community input is welcomed.

Twitter provides an unofficial forum for sharing information about digital preservation resources online, as do many personal collections of bookmarks hosted in social bookmarking tools.

Other file format resources are maintained outside of the digital community, the most comprehensive being Wikipedia. Wotsit.org maintains a similarly impressive array of format information. These appear to have been under-utilized in most digital preservation registry initiatives to date.

4. DRAWBACKS OF CURRENT APPROACHES

4.1 Lack of content

Almost without exception, the tool and format registries provided by the digital preservation community suffer from inadequate amounts of content. This observation seems at odds with the effort that has been devoted to existing registry initiatives where the focus has typically been placed on designing detailed data models

³ OPF Tool registry: <http://wiki.opf-labs.org/display/SPR/Digital+Preservation+Tools>

⁴ AQUA <http://wiki.opf-labs.org/display/AQuA/Home>; SPRUCE <http://wiki.opf-labs.org/display/SPR/Home>.

⁵ Digital Curation Exchange: <http://digitalcurationexchange.org/>

⁶ NDIIPP Partner Tools & Services list:
<http://www.digitalpreservation.gov/tools/>

⁷ Digital Formats Sustainability:
<http://www.digitalpreservation.gov/formats/>

and building systems to manage and publish the resulting RI. The result is theoretically capable replicas and systems, which are largely empty of their most important feature: the data. We suggest that the biggest challenges facing these initiatives are not related to managing or publishing RI, but in capturing and recording it

4.2 Duplication and reinvention

A considerable number of DP community-created web pages list digital preservation tools. Most have some unique entries, though many contain entries duplicated across other entries (albeit with slightly different descriptions). The result is that users are unable to easily find the tools they need and precious DP community resources are spent needlessly reinventing the wheel or aspects of the wheel. For example, more than one institution has developed its own checksum tool for digital preservation purposes.

4.3 Lack of use

It is undeniable that despite the massive investments made to establish representation information registries, the current initiatives are under-utilized. Much effort has been devoted over the past decade to developing new digital preservation tools and approaches, but insufficient attention has been paid to the needs of the users. The result is a mismatch between preservation tools, and user requirements.⁸

This may be down to insufficient understanding about use cases and requirements. RI repository use cases are undeniably unclear, though it may also be a case of chicken and egg: which comes first, the RI, or an understanding of how RI should be used? Perhaps the community still has insufficient detailed understanding of how RI fits into a preservation strategy and the relationship between RI requirements and different preservation strategies. Or is it perhaps a case that we have not yet reached the stage, from a temporal perspective, where we need much more than file format specifications. Whatever the reason, it will only be solved by greater collaboration and engagement with the user community.

5. ADVANTAGES AND DISADVANTAGES OF A COMMUNITY & COLLABORATIVE APPROACH

A community-based approach to collecting and managing representation information has potential to resolve many of the drawbacks in current approaches. For example:

- It is user focused, so the final data is more likely to meet the needs of end users and is therefore more likely to be used.
- It puts the initial focus on capturing content, thereby increasing the flow of incoming data and increasing the chances of reaching that critical mass.
- A single, concerted and collaborative effort will minimize efforts wasted through duplication and reinvention
- The end result is likely to be of a higher quality with less effort from any one participant (and therefore more distributed costs), as it has been refined by the crowd,

⁸ Mashup events have provided a useful forum in which to engage with considerable numbers of users, capture and publish their requirements and explore solutions by utilizing existing open source software).

with a higher number of contributions and expertise from a wider cross section of the community.

The risks of a communal and collaborative approach however, cannot be overlooked:

- There may be difficulty reaching consensus about the level and granularity of RI resources required.
- Without sufficient refinement by a number of contributors, content may be of poor quality.
- Success depends on reaching a critical mass of contributions. If this is not reached, the solution may hold few advantages over other approaches.

Individual organizations that have hosted community discussion forums have typically struggled to reach a critical mass of contribution to make the forums a success. This has been the experience of even those with sizeable and engaged communities such as the Digital Curation Centre, the Digital Preservation Coalition or the Open Planets Foundation. The recent proposal for a digital preservation themed Stack Exchange site seeks input and engagement from across the international digital preservation community. While still requiring further support to reach a functional beta stage at the time of writing, it has been successful in soliciting widespread international support and shows promise for a broad community driven approach. However, it has yet to be seen whether this widespread ‘show of hands’ will translate into active and participatory membership.

Collaborative collection approaches must target content at a level of granularity most likely to be relevant to the majority, in order to engage as broad a swathe of the community as possible. We propose that success at this level is most probable if it is a) simple, b) does not require extensive input from contributors, and c) makes use of existing tools and networks. Our answer to this is CRISP.

6. CRISP: A COMMUNITY APPROACH TO COLLECTING REPRESENTATION INFORMATION

CRISP utilizes the power and wisdom of the crowd to identify and share online resources of representation information, beginning with file format specifications. We have selected format specifications as they are the lowest common denominator of representation information: as previously argued, files of a given format and version share a core RI requirement for the same format specification, regardless of the more extensive environment in which they were produced (the RI for which is more likely to differ across different environments and uses). Access to format specifications is necessary for all preserving institutions. This initiative is therefore broadly relevant and with a clearly defined scope.

CRISP is in the early stages of development. The main objective of the initiative is to address the gaps in collection content currently evident in global format registries managed by the digital preservation community. We will, in essence, get the data. Once we have it, we will store it in a preservation-capable store. We expect to expand our scope to preservation tools in the future, but the initial focus is limited to an achievable and easily defined set of data, namely the format specifications. Our solution has yet to be fully implemented but we are confident that it is sufficiently robust and reliable to serve our needs.

Content will be crowd-sourced via two mechanisms that will make it easy for interested parties to participate. The primary method of submitting information is via an online form, hosted on

the Open Planets Foundation website.⁹ Minimum data requirements have been set purposefully low. The only compulsory fields are a) URL and b) tag(s), though additional fields are available and contributors are encouraged in particular to tag their entries by format to support classification and curation at later stages. Registration is not required prior to nomination. This, alongside a small minimal requirement for input and a simple, straightforward interface, ensures the barriers to participation are as low as possible.

The form links directly to a Google spreadsheet, which is publicly available so participants have access to all nominations and are able to re-use the data if desired. A small number of super-users will be identified to promote the initiative and curate the spreadsheet. De-duplication algorithms will eliminate multiple entries for the same resource whilst maintaining the tags applied by different proposers to ensure broad classification relevance.

The second, more experimental approach is via mentions of the @dpref Twitter account. Tweets to this account will be collated and added to the spreadsheet. We were hoping to use a social bookmarking system like Delicious or Diigo, but we found them to either be unreliable or have too high a barrier to submission. Both also failed to have suitable methods for exporting the curated dataset. A Google spreadsheet offers the functionality and access that is needed.

We propose that the repository element of the equation is served by the existing power of well-established web archiving systems, which will harvest sites listed in the spreadsheet and store them as part of an RI ‘collection’. This will, in the first instance, be undertaken by the UK Web Archive. As the spreadsheet will be publicly available and the contents broadly relevant, we hope that the initiative will be more widely adopted by the global preservation community in the near future and that other web archiving institutions will also avail themselves of the resource. By remaining neutral in terms of ownership, it is anticipated that buy-in across the community will be increased.

We are not the first group to propose use of web archives for collecting representation information. The subject has been raised more than once in the IIPC Digital Preservation Working Group. More recently, the web archiving team at the Library of Congress has begun archiving web pages identified in the Digital Formats Sustainability site. However, web archiving alone will not solve the challenge of resourcing and broad relevance to the community. Crowdsourcing has been used by cultural heritage institutions to meet other objectives in recent years, for example correcting OCR text, and has successfully increased the amount of manpower available to an initiative whilst simultaneously raising awareness of the content and increasing use. There is no reason to believe this approach will be any different.

Our proposal is simple, and we are confident that its simplicity will be the key to its success.

7. ISSUES

The main advantages of our approach stem from its low cost, clearly defined scope, and broad relevance. However, we appreciate that it is not without issues:

- There is the risk that the community will not get on board with the initiative. Without a critical mass of

participants, the initiative will not reach the critical mass of content required.

- Champions and curators are required for sustained community engagement and curation of the data prior to harvest: there are costs associated with this
- Legislative issues may prevent interested web archives from sharing their RI collections publicly, lowering the incentive for input from non-crawling institutions
- An automated solution is required to clearly identify openly licensed content that can be freely republished
- There is a risk associated with using free online tools and services, which may be withdrawn or the data lost with no compensation or backups.

These issues will be managed as the initiative develops.

8. CONCLUSION

CRISP offers a low cost and simple solution to the problem of identifying and collecting essential representation information commonly required by the collecting institutions. The main risk lies in garnering sufficient community engagement to ensure RI sources are nominated. If the community does not buy-in to the proposal, then population of the established representation information repositories will continue at the very slow pace we have seen to date. Similarly, without better community engagement, it will be difficult to clearly identify use cases and encourage use of the repositories. Without this, they will fail to be truly integrated into the preservation solutions currently being developed. CRISP is the first step in solving that problem.

9. ACKNOWLEDGMENTS

Our thanks to colleagues in the web archiving and digital preservation teams at the British Library for their input to this idea.

The SPRUCE project is funded by JISC.

10. REFERENCES

- [1] OAIS standard:
http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683
- [2] Brown, A. 2008 ‘White Paper: Representation Information Registries’ Planets project publication http://www.planets-project.eu/docs/reports/Planets_PC3-D7_RepInformationRegistries.pdf
- [3] See for example the OAIS model and Brown (2008), op cit.
- [4] UDFR <http://www.udfr.org/>; GDFR <http://gdfr.info/>; PRONOM <http://www.nationalarchives.gov.uk/PRONOM/>
- [5] Roberts, B. 2011 ‘A New Registry for Digital Preservation: Conceptual Overview’.
<http://www.openplanetsfoundation.org/new-registry-digital-preservation-conceptual-overview>

⁹ The form is available at
<http://www.openplanetsfoundation.org/testbed/digital-preservation-reference-stack-collection-form>

An ontology-based model for preservation workflows

Michalis Mikelakis

Dept. of Informatics

Athens University of Economic and Business

Athens, Greece

mikelakism@aueb.gr

Christos Papatheodorou

Dept. of Archives and Library Science

Ionian University

Corfu, Greece

papatheodor@ionio.gr

ABSTRACT

This paper aims to propose an ontology for the main digital preservation workflows carried out by an organization or an archival system. The proposed ontology covers the entire preservation life cycle, starting from the ingestion of digital resources and including internal functions, such as system administration and preservation planning policies, and access control. Fifty workflow models have been represented using the ontology, which takes into account the special characteristics and features specified by the international standards, as well as the existing metadata schemas for preservation. The proposed ontology supports the decision making of the collection managers, who design preservation policies and follow practices, by providing a knowledge-based tool able to guide, encode and (re)use their reasoning and choices.

Categories and Subject Descriptors

H.3.7 [Information Storage and Retrieval]: Digital Libraries – standards, systems issues.

General Terms

Design, Documentation, Standardization.

Keywords

Digital Preservation Workflows, Ontology, OAIS Model.

1. INTRODUCTION

Digital preservation has attracted the interest of the scientific community during the last decade since it addresses crucial issues for the future of digital data and information stored in large repositories or published on the World Wide Web. The production of digital data nowadays has grown rapidly and it concerns all aspects of human activity, such as health, science, culture, public functions and political decisions. At the same time, the fast changes in technology have shortened the lifespan of digital objects, which, in contrast to analog ones, have no meaning outside the technical environment that they have been designed for. The danger of information loss is even greater for digitally born objects, where the original information cannot be retrieved from any other source in case of media failure, format or tool obsolescence or loss of metadata.

The systems that have been implemented in the area of digital preservation focus mainly on particular preservation activities such as planning, migration or emulation and follow workflows inspired by OAIS model [6]. Some of them integrate a set of tools trying to provide a preservation framework and support organizations to develop policies and workflows for preserving their own material [3, 8, 9, 12]. However these systems do not

offer a model that expresses explicitly and analytically the workflows they perform in order to (i) guide the user throughout the preservation process and (ii) be potentially reused by other implementations.

This paper proposes an ontology that provides a new conceptualization of the OAIS preservation workflows describing the concepts associated with the structure and form of a digital object as well as the complex relationships involved in the preservation process. The choice of creating an ontology was grounded on the expressive power of such knowledge organization and representation schemes. Moreover, the use of an ontology facilitates information reuse. It could easily be used in its entirety by an organization interested in representing information for digital preservation workflows, or integrated with other internal ontologies of the organization. Furthermore, it can be extended by defining new concepts and relationships or even redefining existing ones in order to fit to one's specific needs. The proposed ontology was developed using OWL, a language for authoring ontologies, which has been endorsed by the World Wide Web Consortium (W3C). The use of a language that has been established as a standard agrees with the concept of long-term preservation and ensures that the model will not become obsolete in the future. Thus the paper exploits semantic web tools to contribute to the systematic aggregation and formal expression of the preservation workflows. Hence the preservation workflows for particular collections and digital objects are represented as instances of a conceptual model and formulate a semantic network. These instances can be retrieved (using SPARQL queries), re-used and interlinked to each other or with other metadata concerning the collections and digital objects.

The next section describes the current standards and tools related to workflow management and used by well known initiatives aiming at the development of digital preservation tools. Section 3 presents the proposed model providing a description of the classes and properties of the developed ontology. Section 4 presents how the ontology is used to represent preservation workflows and provides a detailed example concerning the implementation of a specific preservation workflow model. Section 5 describes the user guidance throughout the preservation process with the utilization of the model and the representation of user interactions with the archival system. In the last section we conclude with summarizing the present work and providing directions for future expansion.

2. BACKGROUND

A workflow is defined as the computerized facilitation or automation of a business process, in whole or part [5]. A workflow is a model of an activity, which is consisted of a set of operations or steps. It defines various objects participating in the

flow of the process, such as documents, roles, information exchanged and tools needed for the completion of each step. Every step is generally described by some basic features, which are input information, output information and transformations made by a person or a machine playing a specific role [4].

Workflow management is a technology that has demonstrated a very large expansion and has been adopted in various industries. Organizations develop and use workflow management systems, which are designed according to their internal processes or adjusted to fit their specific needs. A Workflow Management System is defined as “a system that completely defines, manages and executes workflows through the execution of software whose order of execution is driven by a computer representation of the workflow logic” [5].

The vast spread in the development of workflow management products has lead to the need for a common framework, which will define the basic aspects of a workflow management system and provide standards for the development of systems by different vendors. The Workflow Management Coalition (WfMC¹) is a consortium, comprised of adopters, developers, consultants, analysts, as well as university and research groups, whose purpose is to identify common characteristics among workflow management systems and to define standards for the interoperability of such systems. The WfMC has developed the Workflow Reference Model, in order to define a workflow system and to identify the most important interfaces for the interaction between such systems. Under the scope of the Workflow Reference Model, XML Process Definition Language (XPDL) [13] was defined, which is a format to interchange definitions of business process workflows between different workflow products, including both their structure and semantics. XPDL defines an XML schema for specifying the declarative part of a business process. XPDL is not an executable programming language, but a process design format that visually represents a process definition. Another standard created under the WfMC is Wf-XML, which provides web service operations to invoke and monitor a process that might need a long time to complete, so as to facilitate the communication between a process editing tool and a process execution tool, which may be provided by a different vendor.

The mentioned standards focus mainly on providing a representation of a business process. On the other hand, there are executable languages for representing processes. Business Process Execution Language (BPEL) [7] is one such language, which specifies actions within business processes. BPEL uses an XML-based language and provides the capability of interconnecting with outside systems. Processes in BPEL export and import information by using web service interfaces exclusively. BPEL does not provide a strict protocol and there are no explicit abstractions for people, roles, work items, or inboxes. Instead it is a process-centric model that focuses on the interactions and message exchanges that take place in a process.

Another popular business process management tool is jBPM². jBPM is a flexible Business Process Management Suite which models the business goals by describing the steps that need to be executed to achieve a goal and the order of the steps. It uses a flow chart, where a process is composed of tasks that are

connected with sequence flows. There are a lot of other implementations based on the above models, such as Apache OFBiz Workflow Engine³, Apache Agila⁴, Open Business Engine⁵, wfmOpen⁶ and ActiveBPEL⁷.

A suite of tools created for building and executing workflows is Taverna⁸, a domain-independent workflow management system that uses its own definition language. It provides a graphical designer enabling the addition and deletion of workflow components. Taverna does not provide any data services itself, but it provides access and integration of third party services. The SCAPE project⁹, a recently European founded project on preservation, has chosen Taverna as the tool for representing workflows. Preservation processes are realized as data pipelines and described formally as automated, quality-assured preservation Taverna workflows.

The SCAPE working group continues the efforts of the PLANETS project¹⁰, also co-funded by the European Union, which addresses digital preservation challenges. The project’s goal was to build practical services and tools to ensure long-term access to the digital cultural and scientific assets. In general the project provides a detailed implementation of the preservation functions of an OAIS compliant digital repository. The Planets Functional Model is broken down into three Sub Functions: Preservation Watch, Preservation Planning and Preservation Action [10]. These Sub Functions have been mapped to the functions of the OAIS Reference Model. Especially the Planets Preservation Planning Sub Function is based on the OAIS model to describe the functions and processes of a preservation planning component of a digital repository [11, 12].

The project specifies its own workflow description language and execution engine. A preservation workflow consists of a sequence of invocations of services, where the output parameters of one service are mapped to the input parameters of the next one. Furthermore, the Planets Workflow Execution Engine (WEE) introduces the concept of workflow templates, which are predefined workflow definitions. The user interacts with a set of Web Service interfaces through which he can browse the available templates and choose to instantiate and execute those that meet his specific needs [1].

The proposed approach is designed to cover exclusively and with completeness the needs for representing and manipulating preservation workflows. Therefore it should use a language able to express consistently the semantics of the OAIS Reference Model. An additional requirement would be the subsumption of the information for preservation workflows under the linked data framework. For this purpose OWL was opted for the description of the proposed model.

³ <http://incubator.apache.org/ofbiz/>

⁴ <http://wiki.apache.org/agila/>

⁵ <http://obe.sourceforge.net/>

⁶ <http://wfmopen.sourceforge.net/>

⁷ <http://www.activebpel.org/>

⁸ <http://www.taverna.org.uk/>

⁹ <http://www.scape-project.eu/>

¹⁰ <http://www.planets-project.eu/>

¹ <http://www.wfmc.org/>

² <http://www.jbpm.org/>

3. THE PROPOSED MODEL

As mentioned the design of the model was mainly based on the specifications of the OAIS Reference Model. The entities and the messages exchanged among the different functions specified in the OAIS model were combined into logical sequential steps which constitute the basic workflows. In addition, these workflows were enriched with information provided outside of the OAIS model, especially operations defined within the scope of the Planets project¹¹ [2, 9]. These operations focus on specific functions of the preservation process, such as preservation planning, and provide more details refining the steps of the process.

For the design of the ontology, we used Protégé¹² (version 4.1.) an open-source ontology engineering tool, developed at Stanford University. Protégé has been widely used for ontology development, due to its scalability and extensibility with a large number of plug-ins. The classes and properties of the proposed ontology are described in the next sections, while the whole model is presented in Figure 1.

3.1 Preservation Workflows

The OAIS Reference Model has been established as a fundamental design reference model for an archival system and has been widely adopted as a basis in digital preservation efforts in many areas, such as digital libraries, commercial organizations and government institutions. The OAIS model defines the basic entities and functions required by an organization responsible for the preservation of digital information and its availability to a Designated Community and it provides a minimal set of responsibilities for an archive to be called an OAIS. It consists of six main entities, which are Ingest, Archival Storage, Data Management, Administration, Preservation Planning and Access. Each entity plays a specific role in the preservation process.

The OAIS model also defines specific roles which describe the way that external users interact with an archival system and the way that internal users can manage the broader policy of a system. These roles are referred to as Producer, Consumer and Management. Every user can take specific actions according to the available interfaces. A Producer is the person or system which provides the data products to be preserved. An object submitted to the system must have specific characteristics and meet some minimum requirements in order to be accepted. OAIS makes an extensive description concerning the ways for representing information and the structure of a digital object, as well as the forms that it can take inside and outside the scope of an archival system. Before an information package is accepted, the archival system should make sure that it has the required control and rights to ensure the long-term preservation of the information.

Thus the preservation of a digital object is a complex procedure, which follows specific policies and a general strategy defined by the archive management in agreement with the users. It consists of several steps, each of them operated by a number of internal functions of the archival system. Several functions should cooperate sequentially or in parallel via the exchange of objects for a complete preservation process.

¹¹ <http://www.ifs.tuwien.ac.at/dp/plato/intro.html>

¹² <http://protege.stanford.edu/>

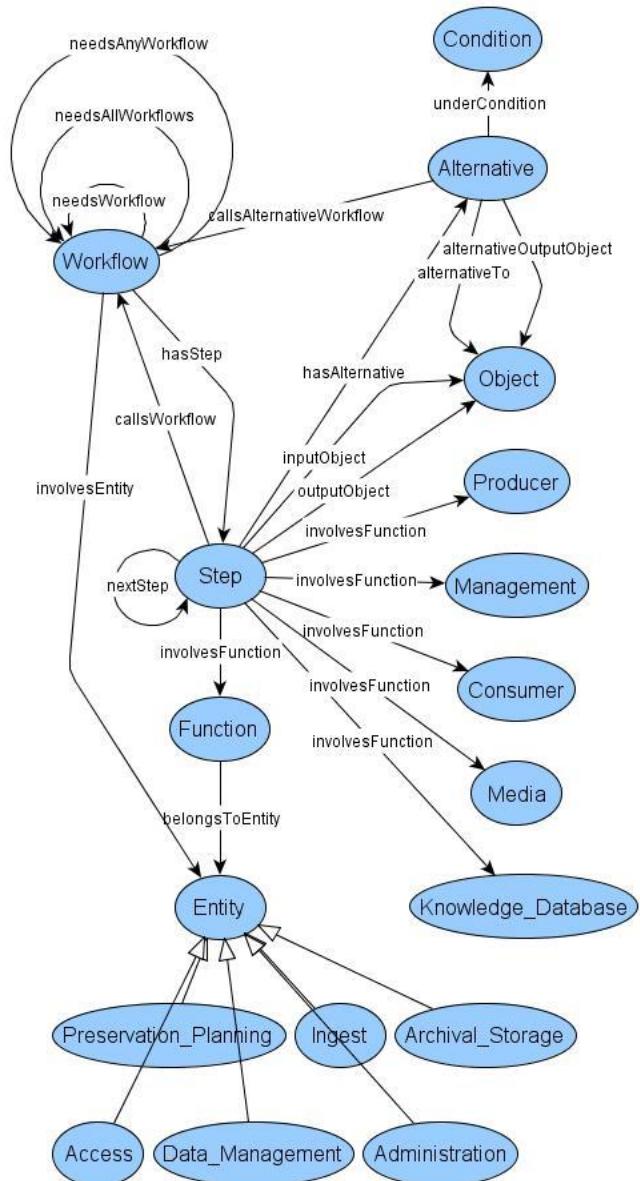


Figure 1. The proposed ontology

An important aspect of an archival system is the way it makes the preserved information available to external users, also referred to as the Designated Community. It should provide a Consumer with search functionalities on metadata kept by the archive or even on the preserved objects themselves. This is accomplished by the iterative submission of queries and the return of query responses.

Based on the above description, some basic concepts that describe the structure of an archival system and the interactions with the users can be concluded. The workflows are divided into six groups, in accordance to the functional entity that is responsible for their execution. Specifically, the workflows are related to Ingest, Archival Storage, Data Management, Administration, Preservation Planning and Access. A workflow may be executed directly and therefore be considered as a primitive workflow, or it may have to wait for another workflow to be completed in order to be able to start. Each workflow consists of one or more steps,

which are executed consecutively or may be executed in parallel. A step has an input and/or output object and is executed by a specific function. After a step is completed it may call the next step(s), call another workflow or end the workflow. The exact classes and properties that constitute the proposed ontology are introduced in the next sections.

3.2 Classes

The classes of the ontology are defined as follows:

Entity: It encompasses the functional entities as described in the OAIS Reference Model. Hence its subclasses are: Ingest, Access, Administration, Archival_Storage, Data_Management and Preservation_Planning.

Function: The entities perform particular functions; according to the OAIS model the subclasses of this class are the following: Activate_Requests, Administer_Database, Archival_Information_Update, Audit_Submission, Co-ordinate_Access_Activities, Co-ordinate_Updates, Customer_Service, Deliver_Response, Develop_Packaging_Designs_and_Migration_Plans, Develop_Preservation_Strategies_and_Standards, Disaster_Recovery, Error_Checking, Establish_Standards_and_Policies, Generate_AIP, Generate_DIP, Generate_Descriptive_Info, Generate_Report, Manage_Storage_Hierarchy, Manage_System_Configuration, Monitor_Designated_Community, Monitor_Technology, Negotiate_Submission_Agreement, Perform_Questions, Physical_Access_Control, Provide_Data, Quality_Assurance, Receive_Data, Receive_Database_Updates, Receive_Submission, Replace_Media.

Role: It includes the main roles of the external entities, as described by the OAIS Reference Model; hence its subclasses are Management, Producer and Consumer.

Object: Every object that may be exchanged between two functions during a digital preservation process. Each object is represented as a subclass of the Object class. According to the OAIS model these subclasses are: AIP, AIP_request, AIP_review, Advice, Alternatives, Appeal, Approved_standards, Assistance, Assistance_request, Audit_report, Bill, Billing_information, Budget, Change_requests, Commands, Cost_estimate, Customer_comments, Customisation_advice, DIP, Data_Formatting_standards, Database_update_request, Database_Update_response, Descriptive_information, Disaster_recovery_policies, Dissemination_request, Documentation_standards, Duplicate_AIP, Emerging_standards, Error_logs, External_data_standards, Final_ingest_report, Inventory_report, Issues, Liens, Migration_goals, Migration_package, New_file_format_alert, Notice_of_data_transfer, Notice_of_shipping_order, Operational_statistics, Order, Payment, Performance_information, Policies, Potential_error_notification, Preservation_requirements, Procedures, Product_technologies, Proposal, Prototype_request, Prototype_results, Quality_assurance_results, Query_request, Query_response, Receipt_confirmation, Recommendations, Report, Report_request, Request_accepted_notification, Request_rejected_notification, Requirements_alerts, Resubmit_request, Review-updates, Risk_analysis_report, SIP, SIP_design, SIP_review, SIP_templates, Schedule_agreement, Security_policies, Service_requirements, Status_of_Updates, Storage_confirmation, Storage_management_policies, Storage_request, Submission_agreement, Survey, System_evolution_

policies, System_updates, Technology_alert, Template, Tools, Unanticipated_SIP_Notification.

Media: According to OAIS this class represents hardware and software settings within the archive.

Workflow: This class is defined as the set of all the preservation workflows. Each entity involves a subset of workflows; the workflows in each entity are modelled as subclasses of the class *Workflow*.

Step: Each workflow consists of a set of distinct steps. The steps of each workflow are modelled as subclasses of the class *Step*.

Alternative: An alternative out of the normal flow in a step, depending on a specific condition, which leads to an alternative output object and may also result in an alternative workflow being called.

Condition: A condition that must be satisfied so as for an alternative to take place. This class is the set of all the conditions that must hold before the execution of alternatives.

Knowledge_Database: The database that stores the gained experience and knowledge from preservation planning activities.

The instances of the mentioned classes correspond to particular functions, steps, workflows, etc. applied by the administrators of a digital repository for the preservation of the objects of particular collections. The ontology provides a rich vocabulary to express in detail and explicitly the actions and the dependencies between them.

3.3 Properties and their constraints

The properties of the ontology correlate its classes defining reasoning paths. The proposed object properties of the ontology are defined as follows:

involvesEntity: This property correlates a workflow to the entity involved in it. Hence the domain of this property is the class *Workflow* and its range the class *Entity*. A constraint is defined on the property imposing that every workflow must be related with exactly one entity.

hasStep: This property denotes that a workflow includes at least one step; it correlates a workflow with all the steps that are needed for the workflow to be completed. Thus the domain of the property is the class *Workflow* and its range is the class *Step*.

involvesFunction: The domain of this property is the class *Step* and its range is the union of the classes *Function*, *Consumer*, *Producer*, *Management*, *Media* and *Knowledge_Database*. Every step must be related to exactly one Function, Consumer, Management, Media, Producer or Knowledge_Database with the property in hand.

belongsToEntity: This property relates a function with the entity it belongs to; thus the domain of the property is the class *Function*, while its range is the class *Entity*. Every function must be related to at least one entity with this property.

inputObject: It defines that a step requires as input an object. Its domain is the class *Step* and its range the class *Object*.

outputObject: It relates a step with an object produced by the step as an output. Its domain is the class *Step* and its range the class *Object*.

nextStep: It correlates a step to all the steps that immediately follow after it. Thus the domain and the range of this property is the class *Step*.

callsWorkflow: It correlates a step with the workflow that is called after its completion, denoting that a workflow might follow a step of a preceding workflow. The domain of the property is the class *Step* and the range the class *Workflow*.

needsWorkflow: It correlates a workflow with the required workflows for its completion. The required workflows must be completed before the beginning of the current workflow. This property has two subproperties, the *needsAllWorkflows* and *needsAnyWorkflow*. The first subproperty means that all the required workflows must be completed before the execution of the workflow in hand and the second subproperty implies that a workflow can begin after the completion of any one of the required workflows.

hasAlternative: Its domain is the class *Step*, while its range is the class *Alternative* and denotes an alternative of a step.

alternativeOutputObject: The property identifies the output object of an alternative step of the given step. Its domain is the class *Alternative* and its range is the class *Object*.

alternativeTo: The domain of this property is the class *Alternative* and its range is the class *Object*. The property defines the output object that has been substituted by the alternative output object (defined by the previous property).

underCondition: The domain of this property is the class *Alternative* and its range is the class *Condition* and denotes that the execution of an alternative step pre-supposes the satisfaction of a condition.

callsAlternativeWorkflow: It denotes that an alternative workflow is called during a step, instead of the workflow that would normally be called. Its domain is the class *Alternative* and its range the class *Workflow*.

Table 1 concludes the ontology object properties along with their constraints. Moreover three datatype properties are introduced that attribute the names and identifiers of the ontology instances, as follows:

workflowId: It is a data type property correlating a workflow with its identifier, which is a unique string. Every workflow must have exactly one identifier.

objectName: It is a data type property correlating an object with a name, which belongs to the string datatype. Every object must have exactly one object name.

stepId: It is a data type property and denotes that every step must have exactly one identifier; thus the domain of the property is the class *Step* and its range the datatype string.

alternativeId: It is a data type property correlating an alternative with its identifier, which is a unique string. Every alternative must have exactly one identifier.

conditionId: It is a data type property correlating a condition with its identifier, which is a unique string. Every condition must have exactly one identifier.

Table 1. The ontology Properties

Name	Domain	Range	Constraints
alternativeOutputObject	Alternative	Object	
alternativeTo	Alternative	Object	
belongsToEntity	Function	Entity	cardinality =1
callsAlternativeWorkflow	Alternative	Workflow	
callsWorkflow	Step	Workflow	
hasAlternative	Step	Alternative	
hasStep	Workflow	Step	min cardinality =1
inputObject	Step	Object	
involvesEntity	Workflow	Entity	cardinality =1
involvesFunction	Step	Function or Consumer or Management or Media or Producer or Knowledge_Database	cardinality =1
needsWorkflow	Workflow	Workflow	
nextStep	Step	Step	Asymmetric, Irreflexive
outputObject	Step	Object	
underCondition	Alternative	Condition	min cardinality =1

4. IMPLEMENTING THE MODEL

The ontology represents each preservation workflow as a subclass of class *Workflow*. It involves exactly one entity that consists of a number of steps, modeled as subclasses of the class *Step*. Totally 50 workflow models have been created covering every possible internal function of an archival system or user interaction with the system incorporated in OAIS. An example that demonstrates the way the proposed ontology reveals explicitly all the characteristics of a preservation workflow, is given in regard to the Ingest entity as follows.

Figure 2 presents the Ingest entity as it is described in the OAIS Functional Model. Ingest provides the services and functions to accept Submission Information Packages (SIPs) from Producers (or from internal elements under Administration control) and prepare the contents for storage and management within the archive [6]. According to Figure 2, the Ingest entity consists of five functions, Receive Submission, Quality Assurance, Generate AIP, Generate Descriptive Info and Co-ordinate Updates. Each function performs specific tasks and exchanges a number of messages and objects. The sequence of the message and object exchanges defines the basic workflows that are specified by the proposed model.

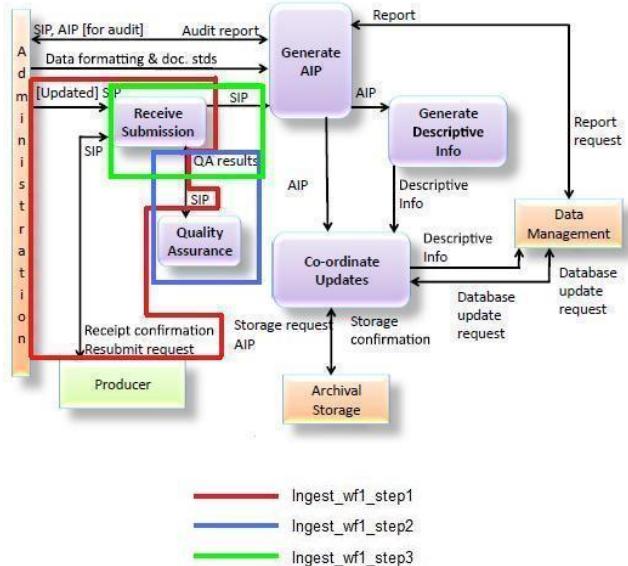


Figure 2. OAIS Ingest Functional Entity

The model decomposes the Ingest entity to four workflows. The first of them is named Ingest_wf1 and consists of three sequentially executed steps, highlighted in Figure 2 as differently coloured frames. Each frame encloses the functions and objects participating in the respective step.

The representation of the workflow Ingest_wf1 by the ontology is shown in Figure 3. The workflow needs any one of the three workflows, namely Ingest_wf4, Administration_wf6 and Administration_wf10, in order to start executing. During the first step, the Receive_Submission function receives a SIP as input from any one of the above workflows and produces a Receipt_Confirmation and a SIP as output to the second step. Alternatively, it can output a Resubmit_request and call the fourth workflow, named Ingest_wf4. During the second step, the Quality_Assurance function receives the SIP, it outputs a Quality_Assurance_results object and continues to the third step, where the Receive_Submission function receives the Quality_Assurance_results as input, outputs a SIP and calls the second workflow named Workflow Ingest_wf2.

An indicative representation of the workflow using the classes and properties of the ontology is shown below. The following fragment from Protégé editor defines that the workflow refers to the Ingest entity and consists of three steps:

```
Ingest_wf1 involvesEntity exactly 1 Ingest
Ingest_wf1 hasStep exactly 1 Ingest_wf1_step1
Ingest_wf1 hasStep exactly 1 Ingest_wf1_step2
Ingest_wf1 hasStep exactly 1 Ingest_wf1_step3
```

The definition of the three steps, encoded in OWL, is given by the following fragment:

```
<SubClassOf>
  <Class IRI="#Ingest_wf1"/>
  <ObjectExactCardinality cardinality="1">
    <ObjectProperty IRI="#involvesEntity"/>
    <Class IRI="#Ingest"/>
  </ObjectExactCardinality>
</SubClassOf>
<SubClassOf>
  <Class IRI="#Ingest_wf1"/>
```

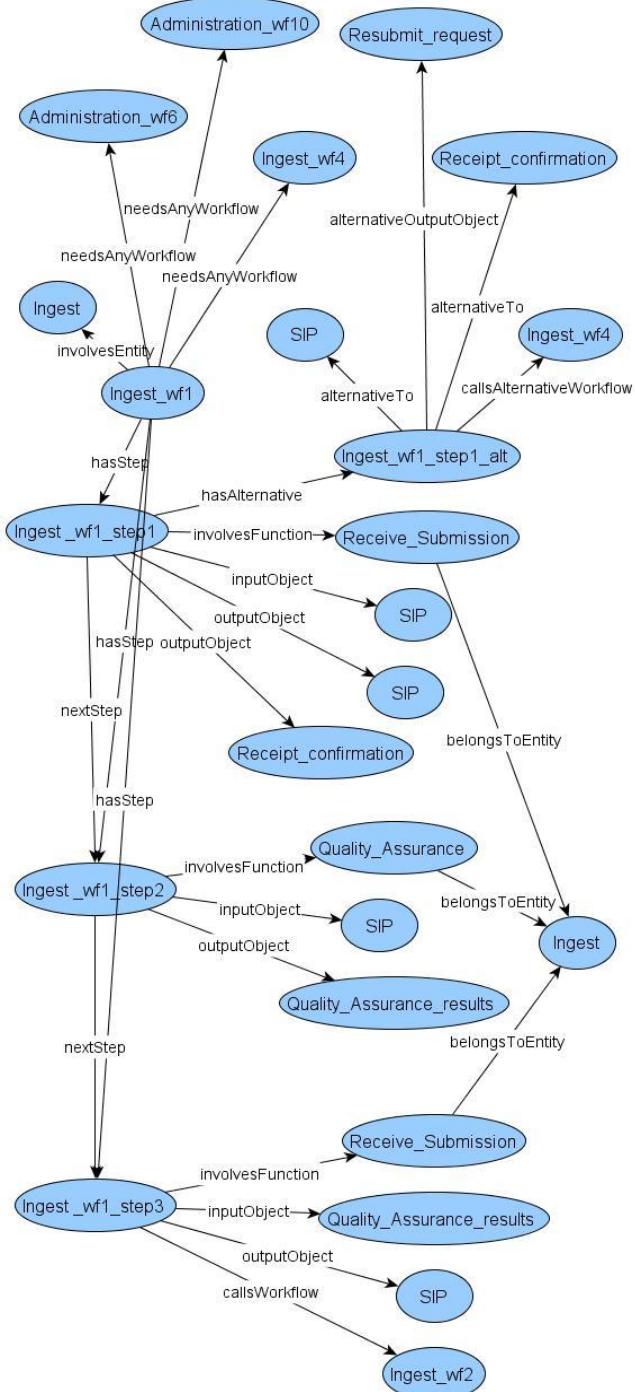


Figure 3. Ingest first workflow

```
<ObjectExactCardinality cardinality="1">
  <ObjectProperty IRI="#hasStep"/>
  <Class IRI="#Ingest_wf1_step1"/>
</ObjectExactCardinality>
</SubClassOf>
<SubClassOf>
  <Class IRI="#Ingest_wf1"/>
  <ObjectExactCardinality cardinality="1">
```

```

<ObjectProperty IRI="#hasStep"/>
<Class IRI="#Ingest_wf1_step2"/>
</ObjectExactCardinality>
</SubClassOf>
<SubClassOf>
  <Class IRI="#Ingest_wf1"/>
  <ObjectExactCardinality cardinality="1">
    <ObjectProperty IRI="#hasStep"/>
    <Class IRI="#Ingest_wf1_step3"/>
  </ObjectExactCardinality>
</SubClassOf>

```

Due to space limits the rest definitions are not given in OWL but as Protégé fragments. The fact that the workflow Ingest_wf1 starts after the completion of any of the workflows Ingest_wf4, Administration_wf6 and Administration_wf10, is declared by the following fragment:

```

Ingest_wf1 needsAnyWorkflow exactly 1 Ingest_wf4
Ingest_wf1   needsAnyWorkflow      exactly 1
Administration_wf6
Ingest_wf1   needsAnyWorkflow      exactly 1
Administration_wf10

```

The definition of the first step of the workflow, named Ingest_wf1_step1, which refers to the subclass Receive_Submission of the class *Function*, as well as its inputs and outputs are presented in the following fragment:

```

Ingest_wf1_step1 involvesFunction exactly 1
Receive_Submission
Ingest_wf1_step1 inputObject exactly 1 SIP
Ingest_wf1_step1 outputObject exactly 1
Receipt_confirmation
Ingest_wf1_step1 outputObject exactly 1 SIP

```

The next step is named Ingest_wf1_step2. However the step Ingest_wf1_step1 has an alternative, named Ingest_wf1_step1_alt. The alternative step produces as output the object named Resubmit_request (instead of a Receipt_confirmation and a SIP) and of course it calls an alternative workflow named Ingest_wf4. These statements are presented in the Protégé fragment:

```

Ingest_wf1_step1 nextStep exactly 1
Ingest_wf1_step2
Ingest_wf1_step1 hasAlternative exactly 1
Ingest_wf1_step1_alt
Ingest_wf1_step1_alt alternativeTo exactly 1
Receipt_confirmation
Ingest_wf1_step1_alt alternativeTo exactly 1 SIP
Ingest_wf1_step1_alt alternativeOutputObject
exactly 1 Resubmit_request
Ingest_wf1_step1_alt callsAlternativeWorkflow
exactly 1 Ingest_wf4

```

The mentioned example constitutes just one indicative case of the set of the encoded workflows that come across during a preservation process. The rest of the workflows are modeled similarly and are available at the URL: <http://www.ionio.gr/~papatheodor/papers/PreservationWorkflows.owl>.

5. GUIDING THE WORKFLOWS

The proposed ontology constitutes a generic model for the representation of preservation workflows. An organization can use the ontology to tailor its own workflows and model its internal structure and functions. The choice of the workflows to be implemented depends on the nature of the organization, its own needs and internal functions as well as the specifications of its archival system. After the selection of the needed workflows, the organization officers should define the instances of the chosen

workflows, their steps, the input and output objects, etc. Given that a subset of the ontology classes have been populated with instances, then a user, who interacts with the archive under a specific role and can execute a number of workflows according to the rights given to this role, could be navigated to the specified paths and monitor the execution of a set of workflows.

The interaction of that user with the archival system can start by selecting the execution of a primitive workflow, i.e. a workflow which is not related to any other workflows through the property *needsWorkflow*. Such a workflow can be executed at any time, regardless of other processes running simultaneously. Then, the user input is combined with information, which is provided to the archive by the prior periodical or on demand execution of other workflows and is stored in the archive database. This information may consist of standards, procedures, templates, statistics or internal policies. The ontology ensures the continuation of the data flows and guides the user by recommending what workflows and steps should be performed at each time point. Moreover, the workflow execution process may ask for the user interaction by providing the user with feedback and requesting additional input.

For instance, a Producer can send a submission information package (SIP) to the Receive_Submission function and call the workflow Ingest_wf1 to accept the SIP and manage the required processing. The person having the role of the producer is modeled as an instance of the class *Producer* and the object provided by the producer is modeled as an instance of the SIP subclass of the class *Object*. The ontology guarantees that the user will follow the processing paths specified by the properties of the ontology and their constraints, presented in Figure 3. The Receive_Submission function receives the SIP provided by the Producer and forwards it to the Quality_Assurance function, while it sends a Receipt_Confirmation object back to the Producer. Alternatively, if there are errors in the submission, a Resubmit_Request is sent back to the Producer and the appropriate workflow is called in order for the proper resubmission of the SIP. Quality_Assurance in turn receives the SIP and send back a Quality_Assurance_Results object. Finally, Receive_Submission, after getting the Quality_Assurance_Results, sends the SIP to the Generate_AIP function and ends workflow Ingest_wf1. The accomplishment of Ingest_wf1 activates the second workflow of the Ingest entity. After the successful performance of a sequence of workflows the object, i.e. the instance of the subclass SIP, is stored in the database of the archival system.

Hence the ontology guides precisely the user to perform the workflows needed to manage the preservation actions for its repository. Concluding, the ontology covers the whole spectrum of the registered workflows and encourages the preservation policy makers and administrators to experiment by either adding new workflow models or by selecting and populating the most appropriate from the existing ones that satisfy the needs of their organization.

6. CONCLUSIONS

Throughout this paper we proposed a model for the representation of the digital preservation workflows, as they can be found in an archival system. Our goal was to cover the entire preservation process and provide a common language to organizations concerned in the field of digital preservation. Therefore the development of the proposed model was mainly based on the OAIS Reference Model. OAIS is a general framework for

understanding and applying concepts needed for long-term digital information preservation. The OAIS Reference Model does not specify a design or implementation. It provides a basis for organizations that aim to implement an archive, by defining general concepts related to long-term preservation. The proposed model provides a tool for specifying the desired preservation activities of an organization as well as it can recommend particular steps and alternatives to a user who runs a preservation activity. Its main advantageous design parameters are the expressiveness to define clearly the preservation workflows, as well as the interoperability and openness ensured by the usage of semantic web languages and open standards.

The present work can be treated in a more detailed way and constitute the basis for a future more elaborated study. The ontology can be used as groundwork for implementing a recommendation system enhanced with a graphical user interface, which will be used by organizations with large volumes of information. Such a system could be fed with a knowledge base, depending on the organization's data and needs, and provide a guide for the entire preservation process.

7. REFERENCES

- [1] ARC. 2009. Guidelines for Creating and Installing IF Preservation Workflows and Templates. IF5-D1. Retrieved June 6, 2012:
http://www.planets-project.eu/docs/reports/Planets_IF5-D1_Creating&Install_IF_Pres_Workflows.pdf.
- [2] Becker, C., Kulovits, H., Rauber, A., and Hofman, H. 2008. Plato: A Service Oriented Decision Support System for Preservation Planning. In *Proceedings of the 8th ACM IEEE Joint Conference on Digital Libraries, JCDL '08*, (Pittsburgh, Pennsylvania, USA, June 16-20, 2008).
- [3] Farquhar, A., and Hockx-Yu, H. 2007. Planets: Integrated Services for Digital Preservation, *The International Journal of Digital Curation, Vol 2 (2)* 88 – 99. Retrieved June 6, 2012:
<http://www.ijdc.net/index.php/ijdc/article/view/45>.
- [4] Fischer, L. 2003. Workflow handbook. Future Strategies Inc., ISBN 0-9703509-4-5.
- [5] Hollingsworth, D. 1995. Workflow Management Coalition. The Workflow Reference Model.
- [6] ISO. Space data and information transfer systems – Open archival information system – Reference model (ISO 14721:2003), 2003.
- [7] OASIS. 2007. Web Services Business Process Execution Language Version 2.0: OASIS Standard. Retrieved June 6, 2012:
<http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.pdf>.
- [8] Rechert, K., Von Suchodoletz, D., Welte, R., Van Den Dobbelaer, M., Roberts, B., Van Der Hoeven, J., Schroder, J. 2009. Novel Workflows for Abstract Handling of Complex Interaction Processes in Digital Preservation. In *Proceedings of the Sixth International Conference on Preservation of Digital Objects, IPRES '09*, (San Francisco, California, October 5-6, 2009).
- [9] Schmidt, R., King, R., Jackson, A., Wilson, C., Steeg, F., and Melms, P. 2010. A Framework for Distributed Preservation Workflows, *The International Journal of Digital Curation, Vol 5 (1)* 205 – 217. Retrieved June 6, 2012:
<http://ijdc.net/index.php/ijdc/article/view/157>.
- [10] Sierman, B., and Wheatley, P. 2010. Evaluation on Preservation Planning with OAIS, based on the Planets Functional Model. PP7-D6.1. Retrieved June 6, 2012:
http://www.planets-project.eu/docs/reports/Planets_PP7-D6_EvaluationOfPPWithinOAIS.pdf.
- [11] 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*, (Vancouver, British Columbia, Canada, June 18-23, 2007), 29-38.
- [12] Strodl, S., and Rauber, A. 2008. Preservation Planning in the OAIS Model. *New Technology of Library and Information Service* (Beijing, China), International Book Trading Corporation, 61 – 68.
- [13] The Workflow Management Coalition Specification, Workflow Management Coalition Workflow Standard: Process Definition Interface -- XML Process Definition Language, Document Number WFMC-TC-1025, 2008.

Interoperability Framework for Persistent Identifiers systems

Barbara Bazzanella
DISI, University of Trento
via Sommarive, 5
38123 Povo Trento, Italy
+39 0461 28 3383
barbara.bazzanella@unitn.it

Emanuele Bellini
FRD, Fondazione Rinascimento Digitale
Via Bufalini 6
50100 Florence, Italy
+39 0555384925
bellini@rinascimento-digitale.it

Paolo Bouquet
DISI, University of Trento
via Sommarive, 5
38123 Povo Trento, Italy
+39 0461 28 3383
bouquet@disi.unitn.it

Chiara Cirinnà
FRD, Fondazione Rinascimento Digitale
Via Bufalini 6
50100 Florence Italy
+39 0555384925
cirinna@rinascimento-digitale.it

Maurizio Lunghi
FRD, Fondazione Rinascimento Digitale
Via Bufalini 6
50100 Florence Italy
+39 0555384925
lunghi@rinascimento-digitale.it

David Giaretta
APA, Alliance for Permanent Access
2 High Street
Yeminster
Dorset DT9 6LF, UK
director@alliancepermanentaccess.org

René van Horik
DANS – Data Archiving and Networked Services
Anna van Saksenlaan 10
2593 HT The Hague, the Netherlands
+31 70 3446484
rene.van.horik@dans.knaw.nl

ABSTRACT

In this paper, we propose an Interoperability Framework (IF) for Persistent Identifiers (PI) systems that addresses functions, roles and responsibilities needed to make heterogeneous PI systems interoperable. The fundamental steps, which provided the main inputs for the design of the model have been: 1) a survey on the use of PI among different stakeholder communities and 2) the definition of interoperability use cases and requirements. The IF is presented as a solution addressing the PI interoperability issues, which have been identified in the survey and have been translated into concrete use cases to serve as requirements for designing the model. Conclusions and intended future work close the paper.

Keywords

Persistent Identifiers (PI), PI Domain (PID), Digital Preservation (DP), Interoperability Framework (IF), reference model, trust.

1. INTRODUCTION

The main goal of this work is to present an Interoperability Framework (IF) for Persistent Identifiers (PI) systems able to overcome the current limits in the use of PI technologies in the actual isolated application domains. When the IF is implemented, the current fragmentation will be reduced, with many additional benefits for the users, provided by some new cross-domain and cross-technology services.

The research work has been carried out through a scientific study and a desk research analysis on the state-of-art of technologies and projects. A questionnaire and some interviews helped to understand the user requirements. The survey investigated current uses and approaches by different user communities of identification systems for digital objects, people, institutions, and few examples of projects trying to implement interoperability among systems. This survey confirmed the absolute lack of such

interoperability and showed that the current systems usually work isolated or in competition, hindering the use of PI across systems and creating complications for the final users. This investigation has been crucial also in order to understand the potential interest by the user communities and the most relevant use cases for our scenario and objectives.

Global and standardized identification systems for people and institutions are not very common. In the digital arena many different systems or methods for objects identification are in use: some of them are valid only locally or for specific types of content, others are used for the identification of physical objects, some are not freely resolvable, others are dynamic and can change over time, and only some of them are really persistent over time and can be considered part of a Digital Preservation (DP) policy. A key concept in this work is the Persistent Identifiers Domain (PID) meaning the system of policy and technology implemented by a user community interested in preserving/using digital contents and managing a PI system for them.

To overcome this fragmented situation, in the framework of the APARSEN Network of Excellence, a reference model has been developed that can be adopted and implemented by any current PI application domain to expose data in a format agreed in the IF, common to all the systems. In this work we ignore all the identification systems not in line with digital preservation criteria and, moreover, we define a benchmark, which specifies the criteria requested to the PI systems to be eligible for our reference model.

2. THE RESEARCH CONTEXT

In order to understand the present work, it is important to contextualize the research within the APARSEN community. Alliance for Permanent Access to the Records of Science in Europe Network (APARSEN), see: <http://www.aparsen.eu> is a Network of Excellence (NoE) co-funded by the European Commission at the call 6 of the FP7, started on the first of January 2011, a consortium of experts on digital preservation with 34 partners in Europe. A NoE is a very specific instrument with the main goal to fight fragmentation of initiatives and research in Europe, a NoE must be thematic and cover a specific topic in line with the FP7 objectives. In Europe even on specific area, like digital preservation, we have a dramatic fragmentation at any level, countries, research centers, professional associations, projects and this causes a waste of resource, investments, impact and competitiveness of our institutions and companies.

APARSEN large consortium brings together a diverse set of practitioner organizations and researchers in order to bring coherence, cohesion and continuity on long-term accessibility and usability of digital information and data researches. The project aims to exploit also this diversity of the partners by building a Virtual Centre of Digital Preservation Excellence. The objective of this project may be simply stated, namely to look across the excellent work in digital preservation which is been carried out in Europe and to try to bring it together under a common vision. The success of the project will be seen in the subsequent coherence and general direction of practices and researches in digital preservation, with an agreed way of evaluating it and the existence of an internationally recognized Virtual Centre of Excellence.

3. PI SURVEY

The main goal of Work Package 22 (WP22) of the APARSEN project is to propose an Interoperability Framework (IF) among different Persistent Identifiers (PI) systems in line with the user communities' requirements. The first year of the WP22 includes two tasks: Task 2210: Survey and benchmarking led by the University of Trento and Task 2220: PI evaluation and integration into an Interoperability Framework and Reference Model led by FRD. The outcome of the Task 2210 and Task 2220 are included in the public deliverable (DE22_1) available at <http://www.aparsen.eu/index.php/aparsen/aparsen-deliverables/>

In order to gain a clearer understanding of the current state of the use of PI systems by different user communities, a questionnaire has been disseminated to the partners belonging to the APARSEN network of excellence and beyond this community (see complete results in Annex I of the DE22_1). The intent of this questionnaire was to explore existing practices, requirements and resources for persistent identification as well as to identify real needs, gaps and challenges related to the use of PI systems. The

questionnaire was spread among several mailing lists such as those hosted by JISC, DPC, APA, DANS, project communities such as Nestor, CASPAR, PLANETS, DPE, PersID, DataCite, etc. and association communities such as AIB, LIBER, CRUI, etc.

Desk research was conducted to identify relevant features, which characterize the main current PI systems and may have an impact on interoperability. This analysis was also useful to understand weaknesses and strengths of each PI system in relation to the user expectations about digital preservation. The results of the desk research activity and the correspondent feature analysis are reported in the Annex II on the DE22_1.

Several APARSEN partners are involved directly in PI projects or services such as STM (DOI), CERN (ORCID), DNB (NBN:DE), DANS (NBN:NL), FRD (NBN:IT), where DOI and NBN are PI systems for digital objects and ORCID is an initiative for PI for authors, or are users of these services, since they manage institutional repositories, usually universities and research institutions, or scientific datasets. Other key players such as DataCite, SURF Foundation, National Library of Sweden, National Library of Australia, National Library of Finland, CrossRef, IETF NBN Working Group have been interviewed during workshops and meetings such as the meeting organized by Knowledge Exchange on "Exploring interoperability of Persistent Object Identifier systems" which produced an important contribution to the identifier interoperability issue through the Den Hague Manifesto <http://www.knowledge-exchange.info/Default.aspx?ID=440>

The point of view and the suggestions of these stakeholders have been taken into account throughout our work.

3.1 Survey structure and Method

In the questionnaire we considered three kinds of persistent identifier systems: 1) PI for digital objects; 2) PI for authors and creators and 3) PI for organizations. The survey was composed of five sections: 1) PI for digital objects; 2) PI for authors/information creators; 3) PI for organizations; 4) Criteria for the adoption of a PI system for digital objects; 5) Digital preservation strategies and practices. In the first three sections we focused on identification practices, limits and requirements for PI for digital objects, authors and institutions. The fourth section contains the criteria adopted by the users for the adoption of PI systems for digital objects, focusing on aspects related to technology, organization of the service, scope, naming rules and expected services. Finally, we addressed issues concerning digital preservation strategies and practices with a special focus on the use of written guidelines, time span, funding and financial sustainability.

3.2 Results

The questionnaire received 103 full responses from participants of three main represented organizations: libraries (47%), universities (27%) and archives (22%)

mainly from academic/research, government and public sectors, 85% of participants were from European countries.

We report here only the results which are more relevant for the design of the IF. The complete analysis of the results is available in the Annex I of the DE22.1.

1) A first analysis was conducted to investigate the **current use of PI systems for digital objects, authors and institutions** among different stakeholder communities. The results show that the DOI (32%), Handle System (28%) and URN (25%) are the most popular PI systems for digital objects even though local identifier systems are still widely adopted (24%). In particular, referring to the stakeholder communities, DOI is the most common system used by universities, libraries, archives and publishers, Handle is mainly adopted by libraries and archives and URN is almost exclusively adopted by libraries. Other systems, like PURL and ARK, are used by a minority of participants (<10%). This scenario shows that PI systems are becoming increasingly oriented towards a specific community, indicating that an IF that allows a cross-community and cross-system communication is clearly needed.

From this result we gained a first indication on which systems have to be considered to be included into the IF. The survey results show also that PI systems for identifying authors are scarcely adopted (52% of participants claimed that they do not use PI for authors). In any case, the IF has to assume the existence of Author ID systems, but avoiding a focus on specific implementations.

A very similar result to the previous one has been found for the persistent identification of organizations. The answers of the participants indicate that there are no specific PI initiatives for organizations. In fact, the majority of the respondents (39%) reported that no system is adopted to identify their organizations. Globally, a fragmentary picture emerges where PI systems adopted for digital objects are slowly adopted for institution. Following the same approach held for author PI systems, the IF assumes the existence of PI systems for organizations avoiding a focus on specific implementations.

2) About the **types of digital objects**, the results of the questionnaire show that textual documents (reported by 98% of participants) and images (selected by 86% of participants) are the most commonly held digital objects. These results suggest that the IF has to address these two types of objects first.

Two other relevant issues deal with **granularity and versioning**. Concerning granularity the survey results show that a finer capability of a PI system to identify and access parts of digital objects is required. Concerning versioning the survey results indicate also that the most common approach for content versioning is linking a new version to the original version through metadata, followed by the practise of considering the new version as an

autonomous object. The use of naming rules is less common among the participants.

Thus the IF should include those PI systems that support the scalability, granularity and versioning issues working mainly at metadata level.

3) One of the objectives of the survey was to investigate the **limits** experienced in using PI systems for digital objects. Some expected results have been reported, such as “locally defined” and “no standard associated” referred to internal identifiers solutions. It is worth mentioning that one of the limits reported regarding DOI and URN is “low adoption” even though these systems are the most widely used systems within our user sample. Finally, “ongoing costs” is one of the most frequently mentioned limits for DOI system.

In general, users perceive a certain level of immaturity for author identification systems which concerns services, trust and authority.

If we compare the obstacles that the respondents reported about the use of PI systems for authors with those about the use of PI systems for organizations, we can notice that the two most frequently selected obstacles are the same: the lack of awareness and the fact that the use of PI systems is not considered a key issue for the organization. This result confirms that one of the main actions of intervention to promote agreement across the different stakeholder communities about the adoption of PI systems should start from increasing the level of awareness about the available systems and their potential positive effects.

4) About **user requirements**, we investigated four domains: technology, organization of the service, scope and naming rules. In terms of technology, our results indicate that users prefer to adopt a system that represents a de facto standard (53%), widely adopted (56%) and based on an open source infrastructure (88%). This was an interesting input in defining the criteria to evaluate as eligible for the IF the PI system (Trusted PI). In terms of the organization of the service, distributed naming authority (48%) and supported by an institution with a mandate (55%) were the preferred options. In terms of scope, the respondents reported to prefer systems open to any digital objects (81%) and cross-community (76%). Finally, concerning naming rules opaque identifiers (55%) (supporting deep granularity (57%)) are preferred above semantic identifiers supporting low-level granularity. No relevant differences were found between the stakeholder groups in the requirements for adopting a PI system for digital objects.

5) The last relevant aspect for the design of the IF deals with **services**. Citability is the most important service associated to the use of PI, followed by services, which support resolution (i.e. global resolution services,

resolution to the resource or to metadata). More than half of the participants reported services for digital object certification among the required services. According to the stakeholders analysis it seems that if citability is a desired service for all the stakeholder groups in long term vision, aspects related to the resolution mechanisms are more relevant for libraries, archives and publishers, while aspects related to certification (and metrics) are more important for universities and research organizations.

Moreover, against our expectations, the PI basic services are those most required. The so-called “advanced services” that were considered most important for the IF received less votes¹. According to this result, the framework design took into account also the objective to empower the basic PI services in addition to set up the conditions for developing new advanced services. This result was crucial in the distinction between different levels of service within the IF infrastructure.

4. USE CASES

Some user scenarios have been defined to introduce and concretize the interoperability concepts and requirements, by providing a number of use cases for IF following the Scenario Based Design technique [6]. We asked the partners to provide one or more scenarios from their experiences about PI use in a long term vision. Since the APARSEN partners are from different domains, the aim was to cover a wide variety of requirements for different stakeholders communities. We have collected 13 scenarios divided in three groups: 1) Scenarios on Citability and Metrics services, 2) Scenarios on Global Resolution Services (GRS) and 3) Scenarios on Digital Object Certification.

These scenarios have been translated into more simple use cases, a schematic framework useful for identifying entities, their relations, functionalities and so forth. The results of this phase have been used as input for the modeling phase.

5. THE PI INTEROPERABILITY FRAMEWORK (IF)

5.1 PI interoperability: related initiatives

Recently, several initiatives and projects have started to address the problem of PI interoperability and solutions have been proposed in different contexts facing some issues at identifier or metadata levels. A first distinction can be made between national and international initiatives. Some initiatives have been emerged within a national context (e.g PILIN² in Australia and RIDIR³ in United

Kingdom) and some of these started as a funded project on a broader geographical level (e.g. PersID⁴). Other initiatives show their presence at an international level (such as ORCID⁵) and aim at introducing global standards for identification, creating a consortium of participating organizations. We can also distinguish between initiatives limited to a specific discipline (e.g. for linguistic resources) or more generic initiatives dealing with a broader range of resources (e.g. OKKAM⁶). Some projects focus exclusively on the problem of PI interoperability for digital objects (e.g. PILIN), while other initiatives address the interoperability issue for author identifiers (e.g. ORCID). The diffusion of a given initiative can also be determined by the way in which the identifiers are assigned by the underlying ID management systems. Some governmental initiatives limit the assignment to people, that embark on an academic career, while other systems allow the registration of any kind of entity (e.g. OKKAM).

5.2 IF definition

Interoperability is an essential feature for federated information architectures which operate in heterogeneous settings also over time. However, the use of the concept is very heterogeneous: interoperability is conceived in an object-related or in a functional perspective, from a user's or an institutional perspective, in terms of multilingualism or of technical means and protocols. Moreover, interoperability is conceived at different levels of abstraction: from the bitstream level up to the semantic interoperability level [1] [2].

In this paper we describe a conceptual framework addressing the identifier interoperability issues, which have been identified in the survey phases and have been translated into concrete scenarios and use cases to serve as requirements for designing the reference model. The IF describes the entities of our domain, their relations and dependencies, the main functionalities and a minimal set of concepts in order to enable the development of specific implementations (i.e. interoperability services).

When the contents from different PIDs (which are currently not interoperable and are completely isolated) are visible through a common interface provided by the IF, users can access and use any content or relation available in the scenario. In particular, we can create any type of service accessing all the contents across the domains and using them even if they are from different PIDs, overcoming in this way a relevant limit in the current situation. The survey on current practices of PI and the

³ Resourcing Identifier Interoperability for Repositories (RIDIR) project
<http://www.jisc.ac.uk/whatwedo/programmes/reppres/ridir.aspx>

⁴ PersID project – <http://www.persid.org/>

⁵ ORCID (Open Researcher and Contributor ID) www.orcid.org/

⁶ OKKAM project <http://www.okkam.org/>

¹ Although the relatively small size of the survey is a concern, there are practical advantages in starting with the basic services.

² PIs Linking Infrastructure (PILIN) project - <http://www.pilin.net.au/>

description of the use cases have been crucial in order to understand the user potential interest and access modalities or specific required functionalities.

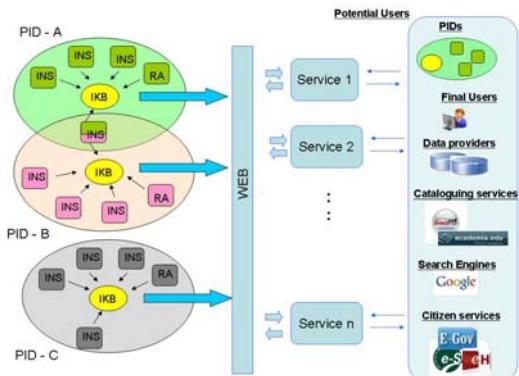


Figure 1 - Interoperability Framework Architecture

5.3 Main assumptions

The IF definition starts from the following main assumptions:

- In the IF we consider only entities identified by at least one PI.
- Only PIDs that meet criteria of Trustworthiness are included in the IF.
- We delegate the responsibility to define relations among the identified entities to the Trusted PIDs.
- We don't address the digital preservation (DP) issues directly but the DP strategy is demanded from the Trusted PIDs. However the IF allows spreading the preservation risk.

According to the main assumptions stated above, only trusted PIDs can join the framework and populate the scenario with their entities. It is important to notice, for the purposes of the present work, that the user community board managing the PID is responsible for guaranteeing suitable policies for any aspect of the DP plan underpinning that system, like for example, the content selection/granularity criteria (included the FRBR⁷ levels), the Trusted Digital Repositories policies and certification, the trustworthiness of the PI management, and so on.

Moreover, within each PID there can be different approaches and architectures to share roles and responsibilities among different components of the system, like the Registration Authority (RA), the Certification Authority (CA), the domain resolver, the digital repository curator and content holders, the DP manager, and so on. The user community is free to choose the best solution and we trust them for the correctness of this choice.

5.4 The reference model

The key actors in the IF are the PI Domains (PIDs) that include in our definition:

- The Registration Agencies (RAs), which manage the allocation and registration of PI according to the trust definition and provide the necessary infrastructure to allow the registrants to declare and maintain the PI-entity relations. We limit to only 3 types of PIDs based on the three different types of identified entities: a) PID for digital objects, b) PID for authors and c) PID for institutions
- The content providers (INS in Figure 1 and 2) that are the institutions responsible for storing, managing and preserving the access to digital contents through the use of PI.
- The resolver is a service able to provide information on the object, its current location and how to get it.

The framework provides a shared conceptual infrastructure to represent the identified entities and their relations within what we call an Interoperability Knowledge Base (IKB), assuming this declared information as guaranteed by trusted PIDs. These relations must be provided by the PIDs when they bring an entity into the interoperability knowledge base. In particular, some trusted PIDs will populate the IKB with their entities presenting these contents following an API so providing specific info requested by the IF. For any digital object the PID, in addition to some descriptive metadata, should declare existing PI (e.g., DOI, NBN), any relation with other objects within the domain and any PI for persons or institutions known by the PID.

In this way, the IKB defines the fundamental relations between the entities in play in the domain (e.g. between objects and PI), creating a layer of accessible knowledge on which interoperability services can be built thanks to the explicit representation of these relations (see Figure 2). Indeed, the knowledge generated independently by the trusted PIDs using the framework, will be exposed on the Web with a common semantics and interface enabling user to access to all the domains and using all the contents even if they are from different PIDs. **Figure 2** shows also that institutions that adopt more than one PI system for their resources, for instance DOI and NBN, contribute to the IKB of the DOI PID and NBN PID with the same relation statements. Thus, IKBs present some overlapping (in Figure 2 is represented by overlapping area between PID-A and PID-B) that can be exploited as a bridge to walk across PIDs and enabling new services to discover new relationships and make inferences on digital resources.

⁷ IFLA- FRBR <http://www.ifla.org/publications/functional-requirements-for-bibliographic-records>

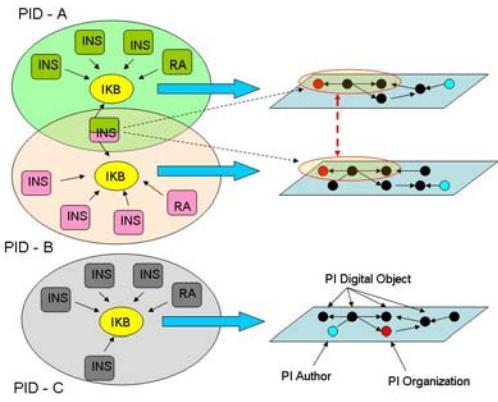


Figure 2 - Significant relations established through the IF across the PID boundaries.

5.5 IF main concepts

Resource: A resource is one of the most primitive concepts in the IF reference model and covers any entity that can be identified by at least one PI. Entities, which are not assigned to a PI, are not eligible for the IF. A resource is a representation of a physical or an abstract entity. Since the concept of resource can be very different in different PIDs, we propose a very general definition, which encompasses the diverse range of digital resources, including resources such as objects, annotations, and metadata. We consider three main kinds of resources in the framework: a) Digital Objects, b) Authors, c) Institutions.

Other kinds of resources can be included in the future with the development of PI systems dealing with other types of entities, such as events, locations and so on.

Digital Object: A digital object is any kind of digital resource, which is identified by at least one PI assigned by a trusted PID. We don't provide a more specific definition because we rely on the definition provided by the trusted PID which has assigned the PI to the resource. A digital object with no PI is not eligible for the IF.

Author: An author is a physical entity, which is the creator of a digital object and is identified by at least one PI assigned by a trusted PID. Whereas digital objects are digital in nature, authors are physical entities which are represented through descriptions (i.e. profiles) in the digital world. Therefore, while a PI for a digital object can point directly to the object, a PI for an author does not point to the author but always to a description of him/her. Moreover the resource, which describes an author, is expected to change as the referent inherently changes across time. Therefore, "the sameness" property of a PI for an author means referring to the "same physical entity" (i.e. the same author and not the same unchanged digital resource), while that of a PI for a digital object means referring to the "same digital entity" (i.e. the same digital resource, in some cases migrated or not, it depends by the PID policy).

Institution: An institution is a physical entity, which affiliates authors and other human agents and is identified by at least one PI assigned by a trusted PID for institutions.

Persistent Identifiers: a PI is a character string used to uniquely identify a resource within a PID regardless of where the resource is located. In the framework we distinguish between 3 kinds of PI.

PID: a PI Domain is a system of users and service providers, which manages the assignment of PI for any type of relevant entities (e.g. digital objects, authors, institutions). Typically, these types of systems are different for different communities and specific for types of objects. PIDs must be trustable in a very long-term vision. We trust PIDs for the implementation of adequate DP rules and strategies.

Policy: the concept represents the set of conditions, rules, restrictions, terms and regulations governing the entire life cycle of a digital resource and its management within a trusted system. This domain is very broad and dynamic by nature. The concept of policy captures the minimal relationships connecting it to the other relevant entities in the framework. The model is extensible and other subclasses of policies could be easily added in future

Resolver: A resolver is a system that provides the link between a PI and information about the object and its current location on Internet, and if available relations with other entities.

User/Actor: An actor is an entity that is external to the interoperability system and interacts with it and uses the related services. Both humans and machine can be users.

5.6 PI trust criteria

In order to design a reliable IF among PI systems, we have to define the criteria that should be met by a PI system. A PI framework has to be reliable to enable the development of advanced services. Thus, only those PIDs that match our criteria for trust will be taken into account as potential component of the framework.

In order to define the trusted PIDs we introduced a small set of criteria distinguishing between mandatory (M) and optional (O) criteria. The criteria are adopted to decide if a PI domain is trusted and eligible for the IF. The definition of these criteria has been suggested by several studies such as, "PI for Cultural Heritage DPE briefing paper" [3], NESTOR reports on trustworthiness of PI systems [4], A Policy Checklist for Enabling Persistence of Identifiers [5], the results of the ERPANET⁸ and DCC⁹ workshops.

⁸ ERPANET workshop Persistent Identifiers Thursday 17th - Friday 18th June 2004-University College Cork, Cork, Ireland www.erpanet.org/events/2004/cork/index.php

⁹ DCC Workshop on Persistent Identifiers 30 June – 1 July 2005 Wolfson Medical Building, University of Glasgow <http://www.dcc.ac.uk/events/pi-2005/>

1. Having at least one Registration Agency (RA).

Within a PI domain it is necessary that a RA is established to assign and maintain the association PI- digital resource. This criterion is considered mandatory in the IF trust assessment.

2. Having one Resolver accessible on the Internet.

To meet this criterion, a resolver able to resolve a PI has to be accessible on the Web. This criterion includes also the capability of a PI to be resolved to an entity represented by a Web page or file, or to both object and metadata or to multiple objects, such as different formats of the same objects, or different content types, through the same PI. We consider this criterion mandatory in our framework.

3. Uniqueness of the assigned PI within the PID.

The RA has to guarantee that a PI is univocally assigned to a digital resource within the PI domain. In fact, since a PI is essentially a string, the uniqueness can be guaranteed only within a domain of reference served by a defined RA. This criterion is considered mandatory in our framework.

4. Guaranteeing persistence of the assigned PI.

Each RA has to guarantee the persistence of the generated PI in terms of preventing the following possible actions:

- a) *String modification*: indicates the PI string update. This kind of updating procedure is not allowed according to our definition of a trusted system.
- b) *Deletion*: indicates the possibility of deleting a PI once it has been created and assigned. This is another process that must be avoided to guarantee trust.
- c) Lack of *sustainability*: indicates that a RA is not able to guarantee its commitment to maintain a PI as far as the identified resource exists. Managing identifiers in a sustainable way is another requisite for a trusted PID. The point a) and b) can be addressed at a functional level of the PI service but they depend by the PID policies; point c) is related to the sustainability of the PI service and the PID business model. This criterion is considered mandatory.

5. User communities, which implement the PID should implement policies for digital preservation (e.g. trusted digital repositories).

It is well known that the main objective of a PI is to provide a reliable access to digital resources in the long term. Thus, if on the one side the RA has to guarantee the persistence of the PI and their association with the identified digital resources (even if they are moved), on the other side, PI should be used to identify stable and preserved digital resources. The content-providers should manage their contents with repositories compliant with standards and common criteria of trustworthiness¹⁰ and

¹⁰ Examples of Trusted digital repository criteria are: *Date Seal of Approval*: <http://www.datasealofapproval.org/>, *Nestor Catalogue of Criteria for Trusted Digital Repositories*: http://files.d-nb.de/nestor/materialien/nestor_mat_08-eng.pdf, *Trusted Digital Repositories: Attributes and Responsibilities*, <http://www.oclc.org/research/activities/past/rfg/trustedrep/repositories.pdf> - *Trustworthy Repositories Audit & Certification*:

implement digital preservation strategies for the resources identified by a PI. This criterion does not require an unlimited guarantee from an organisation but a hand-over procedure should be in place, since content providers manage resources with different life cycles and they can also adopt different commitment to preserve their contents in respect to other institutions.

6. Reliable resolution.

One of the crucial functionalities of a PI system is ensuring that the resolution results of a PI are always the same across time. The definition of the meaning of *the same* is critical, since different domains may manage digital resources at a different level of granularity and require that a PI is generated and assigned to different levels of abstraction of a digital resource. For instance, the PDF version of an article and the HTML version of the same article can be considered "equivalent manifestations" of the same object within the DOI domain (see CrossRef guidelines¹¹), while they would receive two different identifiers in the NBN domain. According to this, the resolution within a PI domain is reliable if the resolution of a PI points to *the same* resource along the time, according to the similarity definition adopted by a PI community. This criterion is considered mandatory.

7. Uncoupling the PI from the resolver.

This criterion is crucial and refers to the PI generation rule defined by a PI system. To be eligible for the IF a PI system has to be based on identifiers whose syntax does not include the URL of the resolver or the content provider in the string. For instance, the NBN syntax definition does not include the URL of the associated NBN resolver. This feature is necessary because the URL of the resolver itself can change. Thus, if a part of the PI string specifies the URL of the resolver domain, all the PI which contain the original URL will become invalid, in case the resolution service is moved to another domain. This criterion is considered mandatory in the proposed IF.

8. Managing the relations between PIs within the domain.

This criterion identifies the possibility to specify the linkage between resources within the PIDs through explicit relations between their identifiers. For example, a PID can make explicit the part-of relation between resources embedding this linkage within the PI string, or using metadata. An example of this kind of relation is that which exists between a resource and the collection of which it is

Criteria and Checklist (TRAC):

<http://wiki.digitalrepositoryauditandcertification.org/pub/Main/ReferenceInputDocuments/trac.pdf>-ISO/DIS 16363:

<http://public.ccsds.org/publications/archive/652x0m1.pdf>,

ISO/DIS 16919

<http://wiki.digitalrepositoryauditandcertification.org/pub/Main/WebHome/RequirementsForBodiesProvidingAuditAndCertification-SecRev1.doc>

¹¹http://www.crossref.org/CrossTech/2010/02/does_a_crossref_do_i_identify_a.html

part. This criterion is considered optional in our framework, but it represents an added value that can speed up the implementation of interoperability services.

We are aware that there are other features and criteria which can be considered in a Trusted PI definition. A critical example is scalability. A PI system that aims to identify an increasing number of objects on Internet (i.e. a global distributed system) must also handle scalability to be considered Trusted. In fact, scalability is one of the basic requirements for the long-term sustainability of every PI service. The main reason why we have not included the scalability as a criterion is due to the variability of the possible technical implementations of a system, and the difficulties in obtaining sufficient information about the technical implementation for making an accurate assessment. The difficulties of obtaining definitive results on such a criterion represent an ongoing concern that has been taken into account in the present work.

6. CONCLUSIONS

In the 2nd year of the APARSEN project the WP22 team will implement a validation mechanism in order to evaluate the Interoperability Framework for PI by around 30 experts, part of them external to the APARSEN consortium. So an action plan to set up a demonstrator for WP22 IF and related services, is under preparation with some external possible synergies with other projects like SCIDIP-ES¹² or other initiatives like ORCID and DOI or NBN large communities. In that demonstrator, some basic services will be tested and refined in order to implement the user requirements collected during the former work in the WP22 with the questionnaire and the use cases definition.

The validation of the model through a user group with experts, including ones external to APARSEN, will be a key strategy to reach consensus and make the model suitable for all the user communities' requirements. Thanks to this consensus building strategy, other user communities beyond the APARSEN consortium will be invited to join the framework and make their content public on the demonstrator, because it is very important to have data from different PIDs and for objects, people and bodies for the potential application spectrum of the user services. By the end of the 2nd year a first prototype with some cross-domains basic services will be set up and become available for the further development of the IF.

7. References

- [1] Stefan Gradmann, INTEROPERABILITY. A key concept for large scale, persistent digital libraries. Digitalpreservationeurope (DPE) project - Briefing Paper - September 2008 <http://www.digitalpreservationeurope.eu/publications/briefs/interooperability.pdf>

[2] Norman Paskin - Interoperability Identifiers - Briefing Paper - Digitalpreservationeurope.eu

<http://www.digitalpreservationeurope.eu/publications/briefs/identifier-interoperability.pdf>

[3] Bellini E, Cirinnà C. and Lunghi, M. PI for Cultural Heritage Digitalpreservationeurope (DPE) EU project – Briefing Paper http://www.digitalpreservationeurope.eu/publications/briefs/persistent_identifiers.pdf

[4] Catalogue of criteria for assessing the trustworthiness of PI systems http://files.dnb.de/nestor/materialien/nestor_mat_13_en.pdf

[5] N. Nicholas, N. Ward and K. Blinco A Policy Checklist for Enabling Persistence of identifiers; D-Lib Magazine Jan/Feb 2009 <http://www.dlib.org/dlib/january09/nicholas/01nicholas.html>

[6] Carroll, J. M. (1995). Introduction: the scenario perspective on system development. In J. M. Carroll (Ed.) *Scenario-based design: envisioning work and technology in system development* (pp. 1-18). New York: John Wiley & Sons, Inc

¹² <http://www.scidip-es.eu>

Conversion and Emulation-aware Dependency Reasoning for Curation Services

Yannis Tzitzikas and Yannis Marketakis and Yannis Kargakis

Institute of Computer Science, FORTH-ICS

Computer Science Department, University of Crete, Greece

{tzitzik|marketak|kargakis}@ics.forth.gr

ABSTRACT

A quite general view of the digital preservation problem and its associated tasks (e.g. intelligibility and task-performability checking, risk detection, identification of missing resources for performing a task) is to approach it from a *dependency management* point of view. In this paper we extend past rule-based approaches for dependency management for modeling also *converters* and *emulators* and we demonstrate how this modeling allows performing the desired reasoning and thus enables offering more advanced digital preservation services. Specifically these services can greatly reduce the human effort required for periodically checking (monitoring) whether a task on a digital object is performable.

1. INTRODUCTION

In digital preservation there is a need for services that help archivists in checking whether the archived digital artifacts remain *intelligible* and *functional*, and in identifying the consequences of probable losses (obsolescence risks). To tackle the aforementioned requirements [14] showed how the needed services can be reduced to *dependency management* services, and how a semantic registry (compatible with OAIS¹) can be used for offering a plethora of curation services. Subsequently, [15] extended that model with *disjunctive dependencies*. The key notions of these works is the notion of *module*, *dependency* and *profile*. In a nutshell, a *module* can be a software/hardware component or even a knowledge base expressed either formally or informally, explicitly or tacitly, that we want to preserve. A module may require the availability of other modules in order to function, be understood or managed. We can denote such *dependency relationships* as $t > t'$ meaning that module t depends on module t' . A *profile* is the set of modules that are assumed to be known (available or intelligible) by a user (or community of users), and this notion allows controlling the number of dependencies that have to be recorded formally (or packaged in the context of an *encapsulation preservation strategy*). Subse-

quently, and since there is not any objective method to specify exactly which are the dependencies of a particular digital object, [10] extended the model with *task-based* dependencies where the notion of task is used for determining the dependencies of an object. That work actually introduced an extensible *object-oriented* modeling of dependency graphs expressed in Semantic Web (SW) languages (RDF/S). Based on that model, a number of services have been defined for checking whether a module is *intelligible* by a community (or for computing the corresponding *intelligibility gap*), or for checking the *performability of a task*. These dependency management services were realized over the available SW query languages. For instance, **GapMgr**² and **PreScan**³ [9] are two systems that have been developed based on this model, and have been applied successfully in the context of the EU project CASPAR⁴. Subsequently, [16] introduced a *rule-based* model which also supports task-based dependencies, and (a) simplifies the disjunctive dependencies of [15], and (b) is more expressive and flexible than [10] as it allows expressing the various properties of dependencies (e.g. transitivity, symmetry) straightforwardly. That work actually reduced the problem of dependency management to *Datalog*-based modeling and query answering.

However, the aforementioned works did not capture *converters* and *emulators*. Since conversion (or migration) and emulation are quite important preservation strategies, a dependency management approach should allow modeling explicitly converters and emulators (and analyze them from a dependency point of view, since they have to be preserved too), and exploit them during the offered preservation services. For example, a sequence of conversions can be enough for vanishing an intelligibility gap, or for allowing performing a task. Since there is a plethora of emulation and migration approaches that concern various layers of a computer system (from hardware to software) or various source/target formats (e.g. see [3] for an overview), it is beneficial to use advanced knowledge management techniques for aiding the exploitation of all possibilities that the existing and emerging emulators/converters enable, and assist *preservation planning* (e.g. [1]). This is crucial since the scale and complexity of information assets and systems evolve towards overwhelming the capability of human archivists and curators (either system administrators, programmers and designers).

¹Open Archival Information System (ISO 14721:2003).

²<http://athena.ics.forth.gr:9090/Applications/GapManager/>

³<http://www.ics.forth.gr/isl/PreScan>

⁴<http://www.casparpreserves.eu/>

In a nutshell, the main contributions of this paper are: (a) we extend the rule-based approach of [16] for modeling explicitly converters and emulators, (b) we demonstrate how this modeling apart from capturing the preservability of converters and emulators, enables the desired reasoning regarding intelligibility gaps, task performability, risk detection etc, (c) we introduce an algorithm for visualizing the intelligibility gaps and thus assisting their treatment, and (d) shows how the approach can be implemented using recently emerged Semantic Web tools. The rest of this paper is organized as follows. Section 2 discusses the motivation and the context of our work. Section 3 introduces the rule based modeling and Section 4 discusses the corresponding inference services. Section 5 shows how the approach can be implemented using Semantic Web tools. Finally Section 6 summarizes, discusses related issues and identifies issues for further research.

2. CONTEXT AND BACKGROUND

Migration (according to Wikipedia) is a set of organized tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another, or from one generation of computer technology to a subsequent generation. The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology. *Emulation* (according to Wikipedia) combines software and hardware to reproduce in all essential characteristics the performance of another computer of a different design, allowing programs or media designed for a particular environment to operate in a different, usually newer environment. Emulation requires the creation of emulators, programs that translate code and instructions from one computing environment so it can be properly executed in another. Popular examples of emulators include QEMU [2], Dioscuri [17], etc. There is currently a rising interest on emulators for the needs of digital preservation [8]. Just indicatively, [18] overviews the emulation strategies for digital preservation and discusses related issues, and several recent projects have focused on the development of emulators for the needs of digital preservation (e.g. see [17] and [11]).

In brief, and from a dependency perspective, we could say that the *migration* process *changes the dependencies* (e.g. the original digital object depends on an old format, while the migrated digital object now depends on a newer format). Regarding *emulation* we could say that the emulation process does not change the dependencies of digital objects. An emulator essentially makes available the behavior of an old module (actually by emulating its behavior). It follows that the availability of an emulator can “satisfy” the dependencies of some digital objects, but we should note that the emulator itself has its own dependencies that have to be preserved to ensure its performability. The same also holds for converters.

Running Example

James has a laptop where he has installed the **NotePad** text editor, the **javac 1.6** compiler for compiling Java programs and **JRE1.5** for running Java programs (bytecodes). He is learning to program in Java and C++ and to this end, and through **NotePad** he has created two files, **HelloWorld.java**

and **HelloWorld.cc**, the first being the source code of a program in java, the second of one in C++. Consider another user, say Helen, who has installed in her laptop the **Vi** editor and **JRE1.5**.

Suppose that we want to preserve these files, i.e. to ensure that in future James and Helen will be able to edit, compile and run these files. In general, to edit a file we need an editor, to compile a program we need a compiler, and to run the bytecodes of a Java program we need a Java Virtual Machine. To ensure preservation we should be able to express the above.

To this end we could use facts and rules. For example, we could state: *A file is editable if it is TextFile and a TextEditor is available*. Since James has two text files (**HelloWorld.java**, **HelloWorld.cc**) and a text editor (**NotePad**), we can conclude that these files are editable by him. By a rule of the form: *If a file is Editable then it is Readable too*, we can also infer that these two files are also readable. We can define more rules in a similar manner to express more task-based dependencies, such as compilability, runability etc. For our running example we could use the following facts and rules:

Facts and Rules	James	Hellen
Facts		
NotePad is a TextEditor	✓	
VI is a TextEditor		✓
HelloWorld.java is a JavaFile	✓	
HelloWorld.cc is a C++File	✓	
javac1.6 is a JavaCompiler	✓	
JRE1.5 is a JVM	✓	✓
gcc is a C++Compiler	✓	
Rules		
A file is Editable if it is a TextFile and a TextEditor is available		
A file is JavaCombilable if it is a JavaFile and a JavaCompiler is available		
A file is C++Combilable if it is a C++File and a C++Compiler is available		
A file is Compilable if it is JavaCombilable or C++Combilable		
A file is a TextFile if it is JavaFile or C++File		
If a file is Editable then it is Readable		

Table 1: Modeling the running examples with Facts and Rules

The last two columns indicate which facts are valid for James and which for Helen. From these we can infer that James is able to compile the file **HelloWorld.java** and that if James sends his TextFiles to Helen then she can only edit them but not compile them since she has no facts about Compilers.

Let us now extend our example with *converters* and *emulators*. Suppose James has also an old source file in Pascal PL, say **game.pas**, and he has found a *converter* from Pascal to C++, say **p2c++**. Further suppose that he has just bought a smart phone running Android OS and he has found an *emulator* of WinOS over Android OS. It should follow that James can run **game.pas** on his mobile phone (by first converting it in C++, then compiling the outcome, and finally by running over the emulator the executable yielded by the compilation). ◇

Regarding curation services, we have identified the following key requirements

Task-Performability Checking. To perform a task we have to

perform other subtasks and to fulfil associated requirements for carrying out these tasks. Therefore, we need to be able to decide whether a task can be performed by examining all the necessary subtasks. For example, we might want to ensure that a file is runnable, editable or compilable. This should also exploit the possibilities offered by the availability of converters. For example, the availability of a converter from Pascal to C++, a compiler of C++ over Windows OS and an emulator of Windows OS over Android OS should allow inferring that the particular Pascal file is runnable over Android OS.

Risk Detection. The loss or removal of a software module could also affect the performability of other tasks that depend on it and thus break a chain of task-based dependencies. Therefore, we need to be able to identify which tasks are affected by such removals.

Identification of missing resources to perform a task. When a task cannot be carried out it is desirable to be able to compute the resources that are missing. For example, if Helen wants to compile the file `HelloWorld.cc`, her system cannot perform this task since there is not any C++Compiler. Helen should be informed that she should install a compiler for C++ to perform this task.

Support of Task Hierarchies. It is desirable to be able to define task-type hierarchies for gaining flexibility and reducing the number of rules that have to be defined.

Properties of Dependencies. Some dependencies are *transitive*, some are not. Therefore we should be able to define the properties of each kind of dependency.

Background: Datalog

Datalog is a query and rule language for deductive databases that syntactically is a subset of Prolog. As we will model our approach in Datalog this section provides some background material (the reader who is already familiar with Datalog can skip this section).

The basic elements of Datalog are: *variables* (denoted by a capital letter), *constants* (numbers or alphanumeric strings), and *predicates* (alphanumeric strings). A *term* is either a constant or a variable. A constant is called *ground term* and the *Herbrand Universe* of a Datalog program is the set of constants occurring in it. An *atom* $p(t_1, \dots, t_n)$ consists of an n -ary predicate symbol p and a list of arguments (t_1, \dots, t_n) such that each t_i is a term. A *literal* is an atom $p(t_1, \dots, t_n)$ or a negated atom $\neg p(t_1, \dots, t_n)$. A *clause* is a finite list of literals, and a *ground clause* is a clause which does not contain any variables. Clauses containing only negative literals are called *negative clauses*, while *positive clauses* are those with only positive literals in it. A *unit clause* is a clause with only one literal. *Horn Clauses* contain at most one positive literal. There are three possible types of Horn clauses, for which additional restrictions apply in Datalog:

- *Facts* are positive unit clauses, which also have to be ground clauses.
- *Rules* are clauses with exactly one positive literal. The positive literal is called the *head*, and the list of negative literals is called the *body* of the rule. In Datalog, rules also must be *safe*, i.e. all variables occurring in

the head also must occur in the body of the rule.

- A *goal clause* is a negative clause which represents a query to the Datalog program to be answered.

In Datalog, the set of predicates is partitioned into two disjoint sets, *EPred* and *IPred*. The elements of *EPred* denote extensionally defined predicates, i.e. predicates whose extensions are given by the facts of the Datalog programs (i.e. tuples of database tables), while the elements of *IPred* denote intensionally defined predicates, where the extension is defined by means of the rules of the Datalog program.

In our context, the proposed implementation is described at Section 5.

3. THE RULE-BASED MODEL

In accordance to [16], digital files and profiles (as well as particular software archives or system settings) are represented by facts (i.e. database tuples), while task-based dependencies (and their properties) are represented as Datalog rules. To assist understanding, Figure 1 depicts the basic notions in the form of a rather informal concept map, in the sense that a rule-based approach cannot be illustrated with a graph in a manner both intuitive and precise.

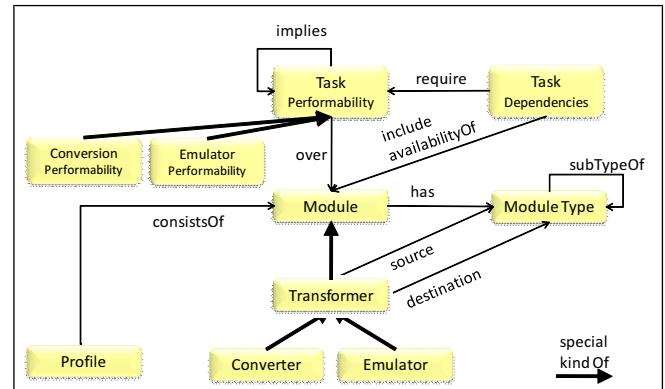


Figure 1: Informal concept map

Digital Files, Type Hierarchies, and Profiles

Digital files and their types are represented as EDB facts using predicates that denote their types, e.g. for the three files of our running example we can have the facts shown in the left column of the following table. Software components are described analogously (e.g. see right column).

Facts	
for digital files	for software components
JavaFile(HelloWorld.java).	TextEditor(vi).
C++File(HelloWorld.cc).	JVM(jre1.5win)
PascalFile(game.pas).	JVM(jre1.6linux)

Each file can be associated with more than one type. In general we could capture several features of the files (apart from types) using predicates (not necessarily unary), e.g. `LastModifDate(HelloWorld.java, 2008-10-18)`.

The types of the digital files can be organized *hierarchically*, and such taxonomies can be represented with rules, e.g. to define that every `JavaFile` is also a `UTF8File` we must add the rule `UTF8File(X) :- JavaFile(X).`

A *profile* is a set of facts, describing the modules available (or assumed to be known) to a user (or community). For example, the profiles of James and Helen are the ticked facts in the corresponding columns of Table 1.

Task-Dependencies and Task Hierarchies

We will also use (IPred) predicates to model tasks and their dependencies. Specifically, for each real world task we define *two* intensional predicates: one (which is usually unary) to denote the (performability of the) task, and another one (with arity greater than one) for denoting the dependencies of the task. For instance, `Compile(HelloWorld.java)` will denote the compilability of `HelloWorld.java`. Since its compilability depends on the availability of a compiler (specifically a compiler for the Java language), we can express this dependency using a rule of the form: `Compile(X) :- Compilable(X, Y)` where the binary predicate `Compilable(X, Y)` is used for expressing the appropriateness of a `Y` for compiling a `X`. For example, `Compilable(HelloWorld.java, javac 1.6)` expresses that `HelloWorld.java` is compilable by `javac 1.6`. It is beneficial to express such relationships at the class level (not at the level of individuals), specifically over the types (and other properties) of the digital objects and software components, i.e. with rules of the form:

```
Compilable(X, Y) :- JavaFile(X), JavaCompiler(Y).
Compilable(X, Y) :- C++File(X), C++Compiler(Y).
Runnable(X, Y)   :- JavaClassFile(X), JVM(Y).
Editable(X, Y)  :- JavaFile(X), TextEditor(Y).
```

Relations of higher arity can be employed based on the requirements, e.g.:

```
Run(X) :- Runnable(X, Y, Z).
Runnable(X, Y, Z) :- JavaFile(X), Compilable(X, Y), JVM(Z)
```

We can express *hierarchies of tasks* as we did for file type hierarchies, for enabling deductions of the form: “if we can do task A then certainly we can do task B”, e.g. “if we can edit something then certainly we can read it too” expressed as : `Read(X) :- Edit(X)`.

We can also express *general properties* of task dependencies, like *transitivity*. For example, from `Runnable(a.class, JVM)` and `Runnable(JVM, Windows)` we might want to infer that `Runnable(a.class, Windows)`. Such inferences can be specified by a rule of the form:

```
Runnable(X, Y) :- Runnable(X, Z), Runnable(Z, Y).
```

As another example, `IntelligibleBy(X, Y) :- IntelligibleBy(X, Z), IntelligibleBy(Z, Y)`. This means that if `X` is intelligible by `Z` and `Z` is intelligible by `Y`, then `X` is intelligible by `Y`. This captures the assumptions of the dependency model described in [14] (i.e. the transitivity of dependencies).

Modeling Converters

Conversions are special kinds of tasks and are modeled differently. In brief to model a converter and a corresponding conversion we have to introduce one unary predicate for modeling the converter (as we did for the types of digital files) and one rule for each conversion that is possible with that converter (specifically one for each supported type-to-type conversion).

In our running example, consider the file `game.pas` (which contains source code in Pascal PL), and the converter `p2c++`

from Pascal to C++. Recall that James has a compiler for C++. It follows that James can compile `game.pas` since he can first convert it in C++ (using the converter), then compile it and finally run it. To capture the above scenario it is enough to introduce a predicate for modeling the converters from Pascal to C++, say `ConverterPascal2C++`, and adding the following rule:

```
C++File(X) :- PascalFile(X), ConverterPascal2C++(Y).
```

Since the profile of James will contain the facts `PascalFile(game.pas)` and `ConverterPascal2C++(p2c++)`, we will infer `C++File(game.pas)`, and subsequently that this file is compilable and runnable.

Finally we should not forget that a converter is itself a module with its own dependencies, and for performing the intended task the converter has to be runnable. Therefore, we have to update the rule as follows:

```
C++File(X) :- PascalFile(X), ConverterPascal2C++(Y),
Run(Y).
```

Modeling Emulators

Emulation is again a special kind of task and is modeled differently. Essentially we want to express the following: (i) If we have a module `X` which is runnable over `Y`, (ii) and an emulator `E` of `Y` over `Z` (`hosting system=Z, target system=Y`), (iii) and we have `Z` and `E`, (iv) then `X` is runnable over `Z`. For example, consider the case where:
`X=a.exe` (a file which is executable in Windows operating system),
`Y=WinOS` (the Windows operating system),
`Z=AndroidOS` (the Android operating system), and
`E=W4A` (i.e. an emulator of WinOS over AndroidOS).

In brief, for each available emulator (between a pair of systems) we can introduce a unary predicate for modeling the emulator (as we did for the types of digital files, as well as for the converters), and writing one rule for the emulation.

For example, suppose we have a file named `a.exe` which is executable over WinOS. For this case we would have written:

```
Run(X)   :- Runnable(X, Y)
Runnable(X, Y) :- WinExecutable(X), WinOS(Y)
```

and the profile of a user that has this file and runs WinOS would contain the facts `WinExecutable(a.exe)` and `WinOS(mycomputer)`, and by putting them together it follows that `Run(a.exe)` holds. Now consider a different user who has the file `a.exe` but runs `AndroidOS`. However suppose that he has the emulator `W4A` (i.e. an emulator of WinOS over AndroidOS). The profile of that user would contain:

```
WinExecutable(a.exe)
AndroidOS(mycomputer) // instead of WinOS(mycomputer)
EmulatorWinAndroid(W4A)
```

To achieve our goal (i.e. to infer that `a.exe` is `runnable`), we have to add one rule for the emulation. We can follow two approaches. The first is to write a rule that concerns the `runnable` predicate, while the second is to write a rule for classifying the system that is equipped with the emulator to the type of the emulated system:

A. Additional rule for Runnable

This relies on adding the following rule:

```
Runnable(X,Y,Z) :- WinExecutable(X),
    EmulatorWinAndroid(Y), AndroidOS(Z)
```

Note that since the profile of the user contains the fact `EmulatorWinAndroid(W4A)` the body of the rule is satisfied (for `X=a.exe`, `Y=W4A`, `Z=myComputer`), i.e. the rule will yield the desired inferred tuple `Runnable(a.exe,W4A,mycomputer)`.

Note that here we added a rule for the `Runnable` which has 3 variables signifying the ternary relationship between executable, emulator and hosting environment.

B. Additional type rule (w.r.t. the emulated Behavior)

An alternative modeling approach is to consider that if a system is equipped with one emulator then it can also operate as the emulated system. In our example this can be expressed by the following rule:

```
WinOS(X) :- AndroidOS(X), EmulatorWinAndroid(Y).
```

It follows that if the profile of the user has an emulator of type `EmulatorWinAndroid` (here `W4A`) and `mycomputer` is of type `AndroidOS`, then that rule will infer that `WinOS(mycomputer)`, implying that the file `a.exe` will be inferred to be `Runnable` due to the basic rule of `Runnable` which is independent of emulators (i.e. due to the rule
`Runnable(X,Y) :- WinExecutable(X), WinOS(Y).`

Both (A and B) approaches require the introduction of a new unary predicate about the corresponding pair of systems, here `EmulatorWinAndroid`. Approach (A) requires introducing a rule for making the predicate `Runnable` “emulator-aware”, while approach (B) requires a rule for classifying the system to the type of the emulated system. Since emulators are modules that can have their own dependencies, they should be runnable in the hosting system. To require their runnability during an emulation we have to update the above rules as follows (notice that last atom in the bodies of the rules):

<code>A' : Runnable(X,Y,Z) :-</code>	<code> B' : WinOS(X) :-</code>
<code> WinExecutable(X),</code>	<code> AndroidOS(X),</code>
<code> EmulatorWinAndroid(Y), </code>	<code> EmulatorWinAndroid(Y),</code>
<code> AndroidOS(Z),</code>	<code> Runnable(Y,X)</code>
<code> Runnable(Y,Z)</code>	<code> </code>

Synopsis To synthesize, methodologically for each real world task we define two intensional predicates: one (which is usually unary) to denote the *performability* of the task, and another one (which is usually binary) for denoting the dependencies of task (e.g. `Readable` and `Read`). To model a *converter* and a corresponding conversion we have to introduce one unary predicate for modeling the converter (as we did for the types of digital files) and one rule for each conversion that is possible with that converter (specifically one for each supported type-to-type conversion). To model an *emulator* (between a pair of systems) we introduce a unary predicate for modeling the emulator and writing one rule for the emulation. Regarding the latter we can either write a rule that concerns the `Runnable` predicate, or write a rule for classifying the system that is equipped with the emulator to the type of the emulated system. Finally, and since converters and emulators are themselves modules, they have their own dependencies, and thus their *performability* and dependencies (actually their runnability) should be modeled too (as in ordinary tasks).

4. REASONING SERVICES

In general, Datalog query answering and methods of logical inference (i.e. deductive and abductive reasoning) are exploited for enabling the required inference services (performability, risk detection, etc). Here we describe how the reasoning services described at Section 2 can be realized in the proposed framework.

Task-Performability. This service aims at answering if a task can be performed by a user/system. It relies on query answering over the Profiles of the user. E.g. to check if `HelloWorld.cc` is compilable we have to check if `HelloWorld.cc` is in the answer of the query `Compile(X)`. As we described earlier, *converters* and *emulators* will be taken into account, meaning that a positive answer may be based on a complex sequence of conversions and emulations. This is the essential benefit from the proposed modeling. Furthermore, classical *automated planning*, e.g. the STRIPS planning method [6], could be applied for returning one of the possible ways to achieve (perform) a task. This is useful in case there are several ways to achieve the task.

Risk-Detection. Suppose that we want to identify the consequences on *editability* after removing a module, say `NotePad`. To do so: (a) we compute the answer of the query `Edit(X)`, let A be the returned set of elements, (b) we delete `NotePad` from the database and we do the same, let B be the returned set of elements⁵, and (c) we compute and return the elements in $A \setminus B$ (they are the ones that will be affected).

Computation of Gaps (Missing Modules). The gap is actually the set of facts that are missing and are needed to perform a task. There can be more than one way to fill a gap due to the disjunctive nature of dependencies since the same predicate can be the head of more than one rules (e.g. the predicate `TextEditor` in the example earlier). One method for informing the curator about the possible ways to fill it is to construct and visualize a *graph* that contains information about only the related facts and rules. We propose a graph which is actually a form of AND-OR graph. The user can specify the desired depth of that graph, or interactively decide to increase the depth gradually. The graph is actually a compact method for presenting the (possibly numerous) ways to fill a gap. The construction of the graph resembles the way *planning algorithms* (in particular backwards search-based planners) operate. The algorithm starts from the goal and shows the corresponding rules for achieving that goal. Those atoms of the rules which have a grounding that belongs to (or can be inferred from) the facts of the profile at hand, are visualized differently (e.g. colored in green, or enclosed in squares) so that the user can discriminate the missing from the available facts. Figure 2 shows some indicative examples. In all cases the goal is a grounded atom, i.e. `A(1)`, however the rules and the recorded facts are different in each case. In case (I) the graph shows that the gap is a grounded atom (i.e. `C(1)`), while in case (II) the graph shows that the gap is a non grounded atom (i.e. `C(var)`). Case (III) demonstrates a case where more than one rules with the same head are involved, and the depth of the graph is greater than one. The graph makes evident that there are two possible ways to fill the gap; according to the first the

⁵In an implementation over Prolog, we could use the *retract* feature to delete a fact from the database.

gap comprises two non grounded atoms (i.e. $D(var)$ and $E(var)$), while according to the second it consists of one non grounded atom (i.e. $D(var)$).

A recursive algorithm for producing such graphs is given (in pseudocode) at Figure 3. The algorithm takes as input a *goal* (an atom grounded or not), a *depth* (a positive integer ≥ 1) and a *prevNode* (the previous node, it is used only for the recursive calls). Initially, the algorithm is called with the goal of the user (which is a grounded atom) plus the desired depth, and an empty (null) prevNode. The algorithm constructs and returns the corresponding tree graph (like those of Figure 2), whose layout can be derived by adopting one of the several hierarchical graph drawing algorithms.

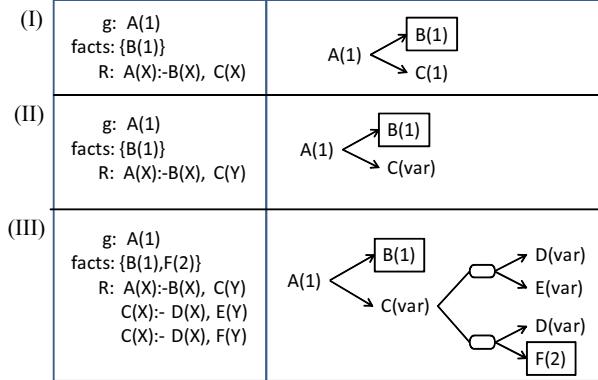


Figure 2: Three examples of gap graphs

Figure 4 shows a small example of a graph of depth equal to 2 where conversion is involved. The graph corresponds to a case where a file `a.pas` is not compilable. The graph makes evident that to turn `a.pas` compilable either a *PascalCompiler* is required or a *Runnable Pascal2Java converter*. Note that if we had a greater depth, then the expansion of *Pascal2Java(var1)* and *Run(var)*, would not necessarily use the same grounding for var1 and var2, although that would be desired. This and other ways to “inject reasoning” in the graph construction is a subject for further research.

Note that the algorithm returns always a tree and it does not do any arbitrary grounding; it is only the original grounded atom (i.e. the original goal) that is propagated based on the rules. Of course if there are rules whose body contain grounded atoms, the latter appear as such in the graph. The algorithm also does not expand a ground atom if inferred.

Complexity. If $|R|$ denotes the number of rules, d the depth, and Q denotes the cost to check whether a fact exists or is inferred (i.e. the cost of query answering), then the time complexity of the algorithm is in $\mathcal{O}(d * Q * |R|)$. Since $|R|$ is usually low, d is an input parameter which again cannot be very big, we can say that the complexity is low.

5. IMPLEMENTATION

There are several possible implementation approaches. Below we describe one Semantic Web-based implementation using RDF/S and *OpenLink Virtuoso* which is a general purpose RDF triple store with extensive SPARQL and RDF support [5]. Its internal storage method is relational, i.e. RDF triples are stored in tables in the form of quads (g, s, p, o) where g represents the graph, s the subject, p the predicate and o the object. We decided to use this system be-

```

Algorithm GapGraph (goal:Atom, depth:Integer, prevNode:Node):Node
(01) If (prevNode=null) then
(02)   gNode = Create node(goal)
(03) else
(04)   gNode = prevNode
(05) hrs = all rules having the predicate of the goal as head
(06) If (|hrs| = 0) then { // the goal predicate is not head in any rule
(07)   headNode = gNode
(08)   return headNode
(09) }
(10) For each hr in hrs
(11)   If (|hrs| > 1) then { // there are > 1 rules having the same head
(12)     ORnode = create node(ORnode)
(13)     create link(gNode→ORnode)
(14)     headNode = ORnode
(15)   } else
(16)     headNode = gNode
(17)   If (IsGrounded(goal)) then { // e.g. consider the goal A(1)
(18)     Ground the corresponding variable in all atoms of the
(19)     body of the rule hr that contain that variable
(20)   }
(21) Let BodyAtoms be the resulting set of body atoms
(22) // if the previous step did not ground anything,
// then BodyAtoms contains the original body atoms
(23)
(24) for each atom in BodyAtoms {
(25)   atomNode = Create node(atom)
(26)   Create link(headNode → atomNode)
(27)   If ((IsGrounded(atom)) and
(28)         (exists in the fact set (or it can be inferred))) then
(29)     Square(atomNode)
}
(30) If (depth > 1) then
(31)   For each atom in BodyAtoms
(32)     If (Square(atomNode)=False) then {
(33)       //atomNode corresponds to atom
(34)       newNode = GapGraph(atom, depth - 1, atomNode)
(35)       Create link(atomNode → newNode)
}
(36)   }
(37) }
(38) Return headNode
  
```

Figure 3: The algorithm that produces gap graphs

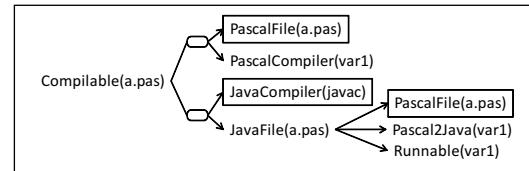


Figure 4: A visualization of a gap graph that involves a converter

cause of its inference capabilities, namely *backward chaining* reasoning, meaning that it does not materialize all inferred facts, but computes them at query level. Its reasoner covers the related entailment rules of `rdfs:subClassOf` and `rdfs:subPropertyOf`, while *user defined custom inference rules* can be expressed using *rule sets*. Practically this means that transitive relations (i.e. `subClassof`, `subPropertyOf`, etc.) are not physically stored in the knowledge base, but they are added to the result set at query answering. *Transitivity* is also supported in two different ways. Given a RDF schema and a rule associated with that schema, the predicates `rdfs:subClassOf` and `rdfs:subPropertyOf` are recognized and the inferred triples are derived when needed. In case of another predicate, the option for transitivity has to be declared in the query.

For our case, we have to “translate” our facts and rules to quads of the form (g, s, p, o) which are actually RDF triples contained in a graph g . The support of different graphs is very useful for the cases of profiles; we can use a different graph for each profile. We will start by showing how facts can be “translated” to RDF quads and later we will show how inference rules can be expressed using ASK and CONSTRUCT or INSERT SPARQL queries. Note that if we use INSERT instead of CONSTRUCT then the new inferred triples will be stored in the triple store (materialization of inferred triples). Hereafter we will use only CONSTRUCT. For better readability of the SPARQL statements below we omit namespace declarations.

Modules: Module types are modeled using RDF classes while the actual modules are instances of these classes. Module type hierarchies can be defined using the `rdfs:subClassof` relationship. For example the fact `JavaFile('HelloWorld.java')` and the rule for defining the module type hierarchy `TextFile(X) :- JavaFile(X)` will be expressed using the following quads:

```
g, <JavaFile>, rdf:type, rdfs:Class
g, <TextFile>, rdf:type, rdfs:Class
g, <JavaFile>, rdfs:subClassof, <TextFile>
g, <HelloWorld.java>, rdf:type, <JavaFile>
```

Profiles: We exploit the availability of graphs to model different profiles, e.g. we can model the profiles of James and Helen (including only some indicative modules), as follows:

```
<jGrph>, <NotePad>, rdf:type, <TextEditor>
<jGrph>, <HelloWorld.java>, rdf:type, <JavaFile>
<jGrph>, <javac_1_6>, rdf:type, <JavaCompiler>
<hGrph>, <VI>, rdf:type, <TextEditor>
<hGrph>, <jre_1_5>, rdf:type, <JavaVirtualMachine>
```

Dependencies: The rules regarding the performability of tasks and their dependencies are transformed to appropriate SPARQL CONSTRUCT statements which produce the required inferred triples. For example, the rule about the compilability of Java files (`Compilable(X,Y) :- JavaFile(X), JavaCompiler(Y)`) is expressed as:

```
CONSTRUCT{?x <compilable> ?y}
WHERE{?x rdf:type <JavaFile>.
      ?y rdf:type <JavaCompiler>}
```

To capture the compilability of other kinds of source files (i.e. C++, pascal etc.) we extend the previous statement

using the UNION keyword (this is in accordance with the Datalog-based rules; multiple rules with the same head have union semantics). For example the case of Java and C++ is captured by:

```
CONSTRUCT{?x <compilable> ?y}
WHERE{
  {?x rdf:type <JavaFile>.
   ?y rdf:type <JavaCompiler>}
  UNION
  {?x rdf:type <C++File>.
   ?y rdf:type <C++Compiler>}
}
```

Finally the unary predicate for the performability of task, here `Compile`, is expressed as:

```
CONSTRUCT{?x rdf:type <Compile>}
WHERE{ {?x <compilable> ?y} }
```

Converters: The rules regarding conversion are modeled analogously, e.g. for the case of a converter from Pascal to C++ we produce:

```
CONSTRUCT{?x rdf:type <C++File>}
WHERE{?x rdf:type <PascalFile>.
      ?y rdf:type <ConverterPascal2C++>.
      ?y rdf:type <Run>}
```

Note the last condition refers is an inferred type triple (Run). If there are more than one converters that change modules to a specific module type then the construct statement is extended using several WHERE clauses separated by UNIONS, as shown previously.

Emulators: Consider the scenario described in section 3, i.e. a user wanting to run a .exe upon his Android operating system. The approach B (which does not require expressing any predicate with three variables), can be expressed by:

```
CONSTRUCT{?x rdf:type <WindowsOS>}
WHERE{?x rdf:type <AndroidOS>.
      ?y rdf:type <EmulatorWin4Android>.
      ?y <runnable> ?x}
```

Services: To realize the reasoning services (e.g. task performability, risk detection, etc), we rely on SPARQL queries. For example to answer if the file `HelloWorld.java` can be compiled we can send the INSERT query about the compilability of the files (as shown previously) and then perform the following ASK query on the entailed triples:

```
ASK{<HelloWorld.java> <compilable> ?y}
```

If this query returns true then there is at least one appropriate module for compiling the file.

The risk-detection service requires SELECT and DELETE SPARQL queries (as discussed at section 4). For example to find those modules whose *editability* will be affected if we remove the module `Notepad`, we have to perform

```
SELECT ?x
WHERE {?x rdf:type <Edit>}

DELETE <Notepad> rdf:type <TextEditor>
```

From the select query we get a set A containing all modules which are editable. Then we remove the triple about

NotePad and perform again the select query, getting a new set B . The set difference $A \setminus B$ will reveal the modules that will be affected. If empty this means that there will be no risk in deleting the **NotePad**.

Based on the above approach we have implemented a prototype system. Its repository containing the facts and rules of the examples of this paper, and behaving as specified by the theory is accessible through a SPARQL endpoint <http://139.91.183.78:8890/sparql>.

6. CONCLUDING REMARKS

In this paper we have extended past rule-based approaches for dependency management for capturing *converters* and *emulators*, and we have demonstrated that the proposed modeling enables the desired reasoning regarding task performability, which in turn can greatly reduce the human effort required for periodically checking or monitoring whether a task on an archived digital object is performable.

We should clarify that we do not focus on modeling, logging or reasoning over *composite tasks* in general (as for example it is done in [4]). We focus on the requirements for ensuring the performability of simple (even atomic) tasks, since this is more aligned with the objectives of long term digital preservation. Neither we focus on modeling or logging the particular workflows or derivation chains of the digital artifacts, e.g. using *provenance* models like OPM or CRM Dig [13]. We focus only the dependencies for carrying out the desired tasks. Obviously this view is less space consuming, e.g. in our running example we do not have to record the particular compiler that was used for the derivation of an executable (and its compilation time), we just care to have at our disposal an appropriate compiler for future use. However, if a detailed model of the process is available, then the dependency model can be considered as a read-only view of that model.

As regards applicability, note that some tasks and their dependencies can be extracted automatically as it has been demonstrated in [9, 7]. As regards available datasets, [12] describes the P2 registry, which uses Semantic Web technologies to combine the content of the *PRONOM Technical Registry*, represented as RDF, with additional facts from DBpedia, currently containing about 44,000 RDF statements about file formats and preservation tools.

In the near future we plan to further elaborate on gap visualization methods, while issues for future research include composite objects (e.g. software components, systems), update requirements, and quality-aware reasoning for enabling quality-aware preservation planning.

Acknowledgements

Work done in the context of NoE APARSEN (Alliance Permanent Access to the Records of Science in Europe, FP7, Proj. No 269977), and SCIDIP-ES (SCience Data Infrastructure for Preservation - Earth Science, FP7).

7. REFERENCES

- [1] C. Becker and A. Rauber. Decision criteria in digital preservation: What to measure and how. *JASIST*, 62(6):1009–1028, 2011.
- [2] F. Bellard. QEMU, a fast and portable dynamic translator. In *Procs of the USENIX Annual Technical Conference*, *FREENIX Track*, pages 41–46, 2005.
- [3] David Giaretta (Editor). *Advanced Digital Preservation*. Springer, 2010.
- [4] D. Elenius, D. Martin, R. Ford, and G. Denker. Reasoning about Resources and Hierarchical Tasks Using OWL and SWRL. In *Procs of the 8th International Semantic Web Conference (ISWC'2009)*, 2009.
- [5] O. Erling and I. Mikhailov. RDF Support in the Virtuoso DBMS. In *Procs of 1st Conference on Social Semantic Web*, 2007.
- [6] R.E. Fikes and N.J. Nilsson. Strips: A new approach to the application of theorem proving to problem solving. *Artificial intelligence*, 2(3-4):189–208, 1972.
- [7] A.N. Jackson. Using automated dependency analysis to generate representation information. In *Procs of the 8th International Conference on Preservation of Digital Objects (iPres'2011)*, 2011.
- [8] B. Lohman, B. Kiers, D. Michel, and van der J. Hoeven. Emulation as a Business Solution: the Emulation Framework. In *Procs of the 8th International Conference on Preservation of Digital Objects (iPres'2011)*, 2011.
- [9] Y. Marketakis, M. Tzanakis, and Y. Tzitzikas. PreScan: Towards Automating the Preservation of Digital Objects. In *Procs of the International Conference on Management of Emergent Digital Ecosystems MEDES'2009*, Lyon, France, October, 2009.
- [10] Y. Marketakis and Y. Tzitzikas. Dependency Management for Digital Preservation using Semantic Web technologies. *International Journal on Digital Libraries*, 10(4), 2009.
- [11] K. Rechert, D. von Suchodoletz, and R. Welte. Emulation based services in digital preservation. In *Procs of the 10th annual joint conference on Digital libraries*, pages 365–368. ACM, 2010.
- [12] D. Tarrant, S. Hitchcock, and L. Carr. Where the Semantic Web and Web 2.0 meet format risk management: P2 registry. In *In Procs of the 6th Intern. Conf. on Preservation of Digital Objects (iPres 2009)*, 2009.
- [13] M. Theodoridou, Y. Tzitzikas, M. Doerr, Y. Marketakis, and V. Melessanakis. Modeling and Querying Provenance by Extending CIDOC CRM. *J. Distributed and Parallel Databases (Special Issue: Provenance in Scientific Databases)*, 2010.
- [14] Y. Tzitzikas. “Dependency Management for the Preservation of Digital Information”. In *Procs of the 18th Intern. Conf. on Database and Expert Systems Applications, DEXA'2007*, Regensburg, Germany, September 2007.
- [15] Y. Tzitzikas and G. Flouris. “Mind the (Intelligibility) Gap”. In *Procs of the 11th European Conference on Research and Advanced Technology for Digital Libraries, ECDL'07*, Budapest, Hungary, September 2007. Springer-Verlag.
- [16] Y. Tzitzikas, Y. Marketakis, and G. Antoniou. Task-based Dependency Management for the Preservation of Digital Objects using Rules. In *Procs of 6th Hellenic Conf. on Artificial Intelligence, SETN-2010*, Athens, Greece, 2010.
- [17] J. Van der Hoeven, B. Lohman, and R. Verdegem. Emulation for digital preservation in practice: The results. *International Journal of Digital Curation*, 2(2), 2008.
- [18] D. von Suchodoletz, K. Rechert, J. van der Hoeven, and J. Schroder. Seven steps for reliable emulation strategies—solved problems and open issues. In *7th Intern. Conf. on Preservation of Digital Objects (iPRES2010)*, pages 19–24, 2010.

Curating the Specificity of Metadata while World Models Evolve

Yannis Tzitzikas and Anastasia Analyti and Mary Kampouraki

Institute of Computer Science, FORTH-ICS

Computer Science Department, University of Crete, Greece

{tzitzik|analyti|mkanbour}@ics.forth.gr

ABSTRACT

The main digital preservation strategies are based on metadata and in many cases Semantic Web languages, like RDF/S, are used for expressing them. However RDF/S schemas or ontologies are not static, but evolve. This evolution usually happens independently of the “metadata” (ontological instance descriptions) which are stored in the various Metadata Repositories (MRs) or Knowledge Bases (KBs). Nevertheless, it is a common practice for a MR/KB to periodically update its ontologies to their latest versions by “migrating” the available instance descriptions to the latest ontology versions. Such migrations incur gaps regarding the specificity of the migrated metadata, i.e. inability to distinguish those descriptions that should be reexamined (for possible specialization as consequence of the migration) from those for which no reexamination is justified. Consequently, there is a need for principles, techniques, and tools for managing the uncertainty incurred by such migrations, specifically techniques for (a) identifying automatically the descriptions that are candidate for specialization, (b) computing, ranking and recommending possible specializations, and (c) flexible interactive techniques for updating the available descriptions (and their candidate specializations), after the user (curator of the repository) accepts/rejects such recommendations. This problem is especially important for curated knowledge bases which have increased quality requirements (as in e-Science). In this paper we elaborate on this problem, we propose a general approach, and discuss examples and a prototype application that we have developed assuming the RFD/S framework.

1. INTRODUCTION

The main (if not all) digital preservation approaches (e.g. the OAIS-based) heavily rely on the existence and curation of metadata, and currently Semantic Web languages, like RDF/S, are increasingly used for expressing them (e.g. see [9, 8]). However ontologies change for various reasons, e.g. an ontology may need to change because it offers a richer conceptualization of the problem domain, the domain of in-

terest has been changed, the perspective under which the domain is viewed has changed, or the user/application needs have changed, and so on.

An important observation is that this evolution happens *independently* of the ontological instance descriptions which are stored in the various Metadata Repositories (MRs) or Knowledge Bases (KBs). With the term *ontological instance description*, (for short “metadata”) we refer to RDF/S [3] descriptions that classify an instance o to a class c or relate two instances o, o' with a property pr . With the term MR or KB, we refer to a stored corpus of ontological instance descriptions, which can be stored in files, in RDF/S databases (i.e. RDF triple-stores [10]), or in the rapidly growing *Linked Open Data* (LOD) cloud [2]. Due to the distributed nature of the Web and the Semantic Web, the evolution of ontologies happens independently of the ontological instance descriptions, e.g. this is the case with ontologies maintained by standardization authorities. However, it is a common practice (mainly for interoperability purposes) for a KB to periodically update its ontologies to their latest versions by “migrating” the stored instance descriptions to the latest ontology versions. This is actually inevitable since scientific terminology and vocabularies constantly evolve. Such migrations are usually not difficult, because newer versions are mainly (or constructed to be) compatible with past ones. Nevertheless, they incur gaps regarding the specificity of the migrated instance descriptions, i.e. inability to distinguish those that should be reexamined (for possible specialization as consequence of the migration) from those for which no reexamination is justified. It follows that quality control is very laborious and error-prone. In this paper we introduce an approach for alleviating this problem.

Consider a corpus of instance descriptions and suppose that at certain points in time we can assert, that the available instance descriptions are the *most specific and detailed descriptions* that are possible with respect to the employed ontology. In other words, our metadata are at a good state. For instance, we can make such an assumption after explicit human (e.g. by the curator of the KB) inspection and verification [4], or in cases where the descriptions have been produced automatically by a method that is guaranteed to produce specific descriptions (e.g. by transforming curated relational data to RDF/S descriptions [14], or by automatic classification to categories each defined by sufficient and necessary conditions, etc.). We will hereafter refer to this assumption by the name *maximum specificity assumption* (for

short *MSA*). It is not hard to see that if the new version of the ontology is *richer* than the past one, then the corpus of the migrated instance descriptions *may no longer satisfy the MSA with respect to the new ontology*.

The ability to identify the instance descriptions that satisfy the *MSA* and those that do not, is useful in order to address questions of the form: (a) for what descriptions can we make the *MSA*? (b) what (class or property) instances should probably be reclassified (to more refined classes or properties), and (c) which are the candidate new classes or properties (refinements) of such instances? The above questions are very useful for curating a corpus of instance descriptions, i.e. for managing its specificity as the corpus (and its ontologies) evolves over time. Without special support, such tasks would be unacceptably expensive and vulnerable to omissions, for large datasets.

The problem occurs in various domains, including Digital Libraries (e.g. as the *Library of Congress Subject Headings LCSH* evolves), in Biomedicine/Bioinformatics (*Gene Ontology*), in e-Government (*oeGOV Ontologies*), etc. Figure 1 sketches some small and indicative examples of ontology evolution. Our work can aid the curation of structured knowledge, i.e. of digital content that is structured according to a structurally object-oriented model, like RDF/S. For instance, the datasets published in LOD fall into this category. For other kinds of content (e.g. documents, audiovisual material, etc), our work can aid the curation of their metadata. For instance consider the *Dublin Core*¹ metadata schema. In many of its elements (attributes) it is suggested to use values coming from controlled (but evolving over time) vocabularies. For instance, this is the case for the attributes *subject* (for describing the topic of the resource), *language* (where it is recommended to use a controlled vocabulary), *coverage* (for describing the spatial or temporal topic of the resource), and *format* (where the use of MIME types are suggested). Furthermore, various *subproperties* for the metadata element *relation* have been proposed in various contexts². As another case, consider annotations/tags of images (e.g. medical images) or entire datasets using elements from an evolving (e.g. medical) ontology, or provenance metadata (e.g. provenance trails of 3D models) that involve artifacts (e.g. photos) and actors (e.g. photo cameras) identified by URIs and described by various metadata from evolving ontologies. Also note that CIDOC CRM which is an ISO standard for the cultural domain, consists of 86 classes and 137 properties, while its extension for digital objects, CRMdig [15], currently contains 31 classes and 70 properties. In general we can say that RDF/S is currently the “lingua franca” for metadata representation and exchange, and this is the reason why in this work we use it as representation framework. Furthermore our work could be used in cases where the *information object* of an information carrier (of any kind), as described in [6], is expressed using RDF/S.

We will explain the main idea of our approach using a small example.

EXAMPLE 1. Consider an e-commerce portal that sells various kinds of products, and suppose the metadata that are

¹<http://dublincore.org/documents/dces/>

²E.g. at EDM (Europeana Data Model).

shown in the left part of Figure 2. Suppose a car *c1* that has been classified under the class *Car*, and a person *p1* that has been classified under the class *Person*, defined in an ontology *Ont1*, and suppose that both classes have no subclasses. Assume that for the current set of instance descriptions according to *Ont1* the *MSA* holds (i.e. they are complete with respect to specificity). We can infer, from this knowledge, that *c1* is not a *Person* and *p1* is not a *Car*. Let *Ont2* be a new version of that ontology which, among other, defines the subclasses of the classes *Car* and *Person*, shown at Figure 2 (right). All subclasses of *Car* are possible classes for *c1*. *Adult* is not a possible class for *c1*, since *c1* was not a person according to *Ont1*. Analogously, none of the subclasses of *Car* is a possible class for *p1*, since *p1* was not a car according to *Ont1*. Moreover, notice that *Ont1* defines a property *owns* and suppose that (*p1* *owns* *c1*) is an instance description. Also notice that *Ont2* defines a subproperty *sells* of *owns* between *Person* and *Car*. This property will be prompted as a possible specialization of the association between *p1* and *c1*.

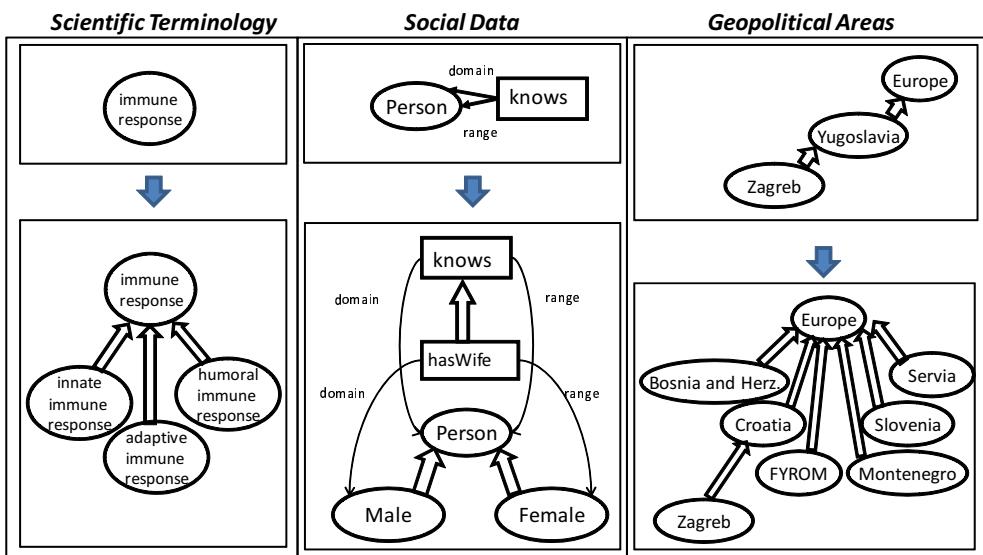
The computation of possible refinements in the general case can be complex since we can have conflicts among (a) new positive knowledge inferable from the instance descriptions and the new schema, (b) new “negative” information inferable from the past negative instance descriptions and the new schema, and (c) the previously computed possible instance descriptions (possible refinements). In fact, our approach resolves such conflicts by considering that (a) has higher priority than (b), and (b) has higher priority than (c). In addition, it should be possible to update correctly the set of possibilities, at scenarios with several successive instance migrations interwoven with several (positive or negative) user feedbacks. Finally, another challenge is to reduce the information that has to be kept to support this scenario, specifically to avoid having to keep negative information of any kind, and to devise compact representations for the possibilities.

We could say that from a more general perspective, our work contributes in enriching the lifecycle of Semantic Web data with quality management, appropriate for scenarios where ontologies evolve frequently and independently from instance descriptions. As a consequence, this allows adopting iterative and agile ontology modeling approaches, appropriate for open environments like Linked Open Data. Note that though there are several works and approaches for dealing with the validity of data during migration in the context of RDF/S (e.g. [11, 7, 13]), there is no work for managing their specificity and quality while ontologies evolve.

The rest of this paper is organized as follows. Section 2 proposes a process model for managing the specificity of metadata, and discusses (mainly through examples) the principles of our approach. Section 3 describes the prototype application that we have developed which is publicly available. Finally, Section 4 concludes the paper and identifies issues for further research.

A thorough elaboration of the problem (that includes formal definitions, algorithms, complexity and experimental results) is available at the technical report

<http://www.ics.forth.gr/~analyti/ipres2012Extended.pdf>.



Cultural Documentation

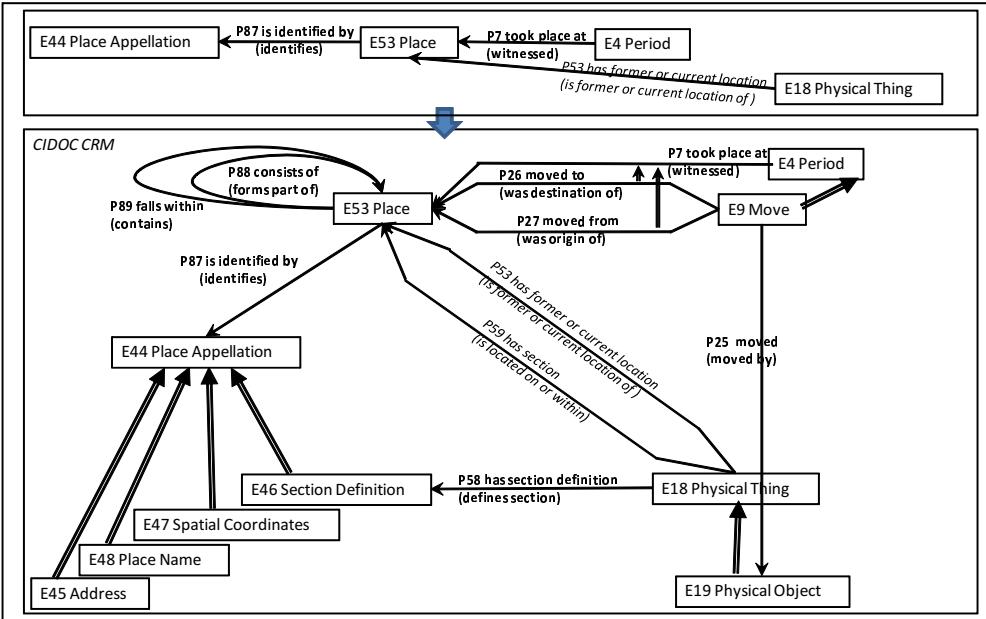


Figure 1: Examples of ontology evolution from various application domains

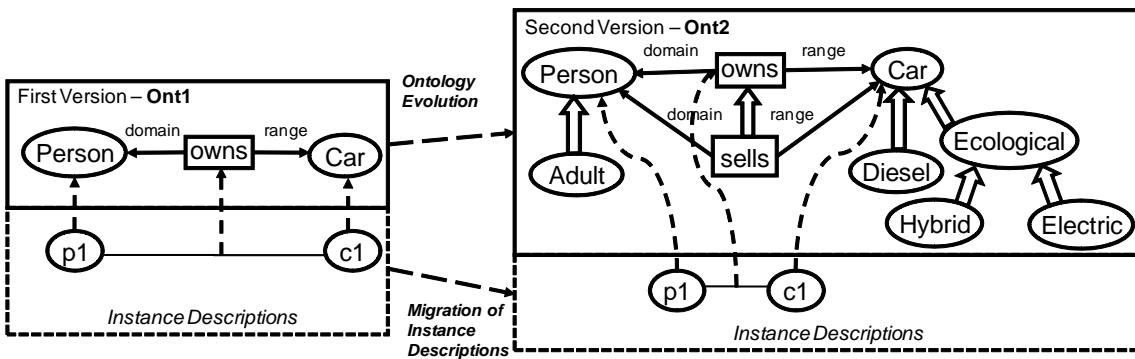
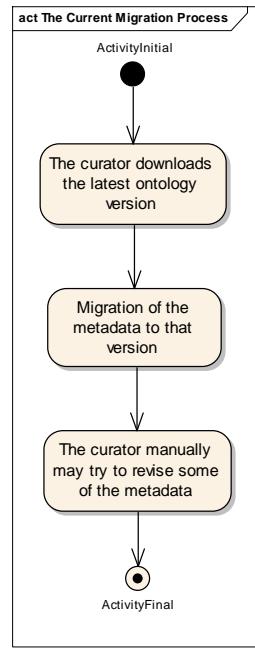


Figure 2: Introductory example

The Current Migration Process



The Proposed Migration Process

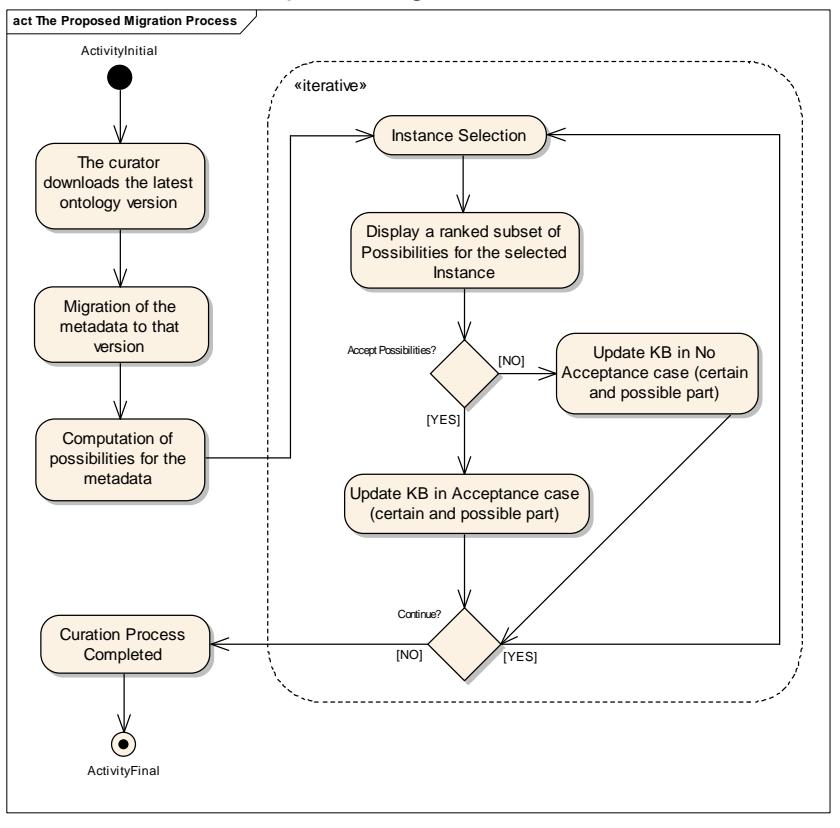


Figure 3: Current and Proposed Migration Process

2. THE APPROACH

2.1 The Life-Cycle

Apart from identifying the information that could be further specialized (as we discussed just before), we would like to aid making it as specific as possible. Therefore, we should support flexible and *interactive* processes for managing the computed possibilities, where the user will be able to either *accept* or *reject* the computed *recommendations*, and eventually update the knowledge base reaching to a state where the *MSA* holds (at least for those resources for which this is possible). The *ranking* of possibilities is important for designing user-friendly interaction schemes, since we may have a high number of recommendations. Essentially, we propose a process like the one sketched in the right part of Figure 3. Specifically, assume that the user selects some instances then the system displays ranked all or some of the possible instance descriptions for the selected instances. The user accepts or rejects these instance descriptions and the system updates appropriately the KB and its possible part. Note that the possible part of the KB is stored explicitly and separately. In our toy example, this means that we can rank the possible classes for *c1*, so that if the user is prompted to select a possible class for *c1*, then *Diesel* and *Ecological* will be the first classes to be displayed. If the user rejects the class *Ecological*, then all its subclasses will be rejected from the possible classes (and this reduces the effort required for reaching a state where the *MSA* holds).

2.2 Foundations and Examples

For reasons of space here we describe only the main points of the theory (the reader can refer to the technical report for the details) and provide some indicative examples.

For expressing (actually bounding) the uncertainty regarding the specificity of a description caused by its migration to a new schema, we introduce the notion of *possible instance triples*. To capture the various application-specific assumptions about the specificity of the descriptions of a KB, we introduce the notion of TFP-partition (True-False-Possible partition). We denote the TFP-partition of a KB K by a triple (of the form $(C_i(K), M_K, P_K)$, the first being a set of positive instance triples (explicitly stated or inferable), the second is a set of negative instance triples, and the last is a set of *possible* instance triples.

We view the *migration* of a set of instance triples to a new schema S' as a *transition* between two TFP-partitions, i.e. $(C_i(K), M_K, P_K) \rightsquigarrow (C_i(K'), M_{K'}, P_{K'})$. Note that the new schema S' can be *backwards* or *non-backwards compatible* with the current schema S . Schema S' is *backwards compatible* with S , if the closure of S (based on the standard inference rules of RDF/S) is subset of the closure of S' .

The transition between two TFP-partitions, is governed by few *postulates* which are very general (i.e. RDF/S independent). We adopt two postulates for the case of backwards compatible, and an additional one (third) for the case of

non-backwards compatible schema evolution.

Specifically the first postulate (P1) gives priority to the positive knowledge inferable from the instance triples and the new schema, and it is consistent with (and reminiscent of) the principle “Recent knowledge prevails the old one” (also, called “Principle of Success” [1] and “Primacy of New Information” [5]).

The second postulate (P2) states that past negative information cannot become possible, meaning that past negative information is preserved as long as it does not contradict with the new positive knowledge.

The last postulate (P3), which is needed only when the new schema is not backwards compatible with the old schema, states and those instance triples that were previously positive, but according to the new schema are not, should be considered in the new TFP-partition as negative (not possible).

Based on the above postulates, a small set of *derivation rules* are defined for carrying out a transition for the case of RDF/S. It is important to note that transitions between TFP-partitions can be defined without having to keep any negative information (i.e. the “M” part of a TFP-partition). Instead only the certain and the possible part of the KB has to be kept, reaching to what we call *extended KB (eKB)*. A further compression of the possible part of the eKB is feasible and suitable for large data sets. Specifically a compact (interval-based) representation of the set of possible instance triples is possible. However the important point is that if the curation process is followed and the curator accepts/rejects the migration-related uncertainties, then the possible part of the KB becomes empty, i.e. no extra storage is required.

Figure 4 illustrates two migrations. The initial schema (at left) contains only one class **Person**. The KB contains only one instance triple, stating that **John** is a **Person**. In the second schema (at the middle) we can see that the class **Person** has been extended with five subclasses. During the migration all these classes are considered as possible classes for **John**. In the figure they are enclosed by a dashed rectangle and the natural numbers indicate their ranking. Now suppose that the system suggests as possible classes for **John** only those with rank 1, i.e. the class **Student** and the class **Employee**. If we suppose that the curator rejects them, then at the right of the figure we can see the new KB. Notice that the set of possible instance triples becomes empty.

Figure 5 illustrates a variation of the previous scenario, where we assume that the system suggests to the curator only three (of the five) possible classes for **John**, namely the classes **Student**, **PostGraduate**, and **Employee**. Here we assume that the curator decides to accept the recommendation **PostGraduate**. At the right diagram we can see the new state of the KB. The set of new possible instance triples contain only that **John** could be **PhD_Student**.

Figure 6 shows an example of a migration to a non backwards compatible schema (notice that one subclassOf relationship has been deleted). The left diagram shows the possible classes for **John** (result of past migrations). At

the bottom of the figure we can see the TFP-partition of these KBs. Note that the previously possible instance triple (**John**, type, **Full-time Permanent Employee**) has been removed and does not belong to $P_{K'}$ because the class **University Employee** is no longer subclass of **Permanent Employee**, and thus **John**, is not an **Permanent Employee**.

The previous examples involved only classes. Properties are analogously treated. An example is shown at Fig 7.

For reasons of completeness, here we describe the rules that determine how the possibilities after a migration are defined. Suppose we are in the context of a transition $(\mathcal{C}_i(K), M_K, P_K) \rightsquigarrow (\mathcal{C}_i(K'), M_{K'}, P_{K'})$. It follows from the postulates, that for a new class c' (i.e. a class that was not element of S), it holds that: (o type c') should be placed at $P_{K'}$ iff:

- (i) $(o$ type $c') \notin \mathcal{C}_i(K')$, and
- (ii) for all not new (i.e. in S) classes c that are superclasses of c' it holds $(o$ type $c) \in (\mathcal{C}_i(K') \cup P_K)$.

Analogously, for a new property pr' (i.e. a property that was not element of S), it holds that: the triple $(o$ pr' $o')$ should be placed at $P_{K'}$ iff:

- (i) the triple $(o$ pr' $o')$ is valid to add, i.e. it respects the domain and range constraints,
- (ii) $(o$ pr' $o') \notin \mathcal{C}_i(K')$, and
- (iii) for all not new (i.e. in S) properties pr that are superproperties of pr' , it holds $(o$ pr $o') \in (\mathcal{C}_i(K') \cup P_K)$.

Regarding deletions, $P_{K'}$ will not contain the instance triples of P_K that their “supertriples” involving old classes or old properties do not belong to $\mathcal{C}_i(K') \cup P_K$. The rest of the instance triples in P_K are transferred to $P_{K'}$.

3. THE PROTOTYPE

We have implemented a proof-of-concept prototype, called **RIMQA (RDF Instance Migration Quality Assistant)**³, supporting the entire lifecycle process. Some screendumps are shown at Figure 8.

The user selects the source ontology (.rdfs file) and a file that contains instance descriptions (.rdf file) with respect to that ontology. The latter file could be the result of applying an export operation over the system that manages the metadata of an archive. Subsequently, the user selects the destination ontology (.rdfs file), which is a subsequent version of that ontology and optionally the user selects a file with possible instance descriptions (.rdf⁴ file) derived from a previous migration with respect to the source ontology and one of its previous versions. The system then automatically migrates the instance descriptions from the source to the destination ontology. Then, it computes the possible instance triples.

After that, if the user presses the “Start Curation” button, the curation process starts. If the user selects the “Statistics” menu, he can see the most indicative statistics about the source and the destination ontology, i.e. (a) the number of original classes, properties, (explicit) schema triples, and instance triples in both ontologies, and (b) the number of added classes and properties, and the number of added and

³The tool is available at <http://www.ics.forth.gr/isl/RIMQA/>.

⁴Note that we use the RDF format in order to store possibilities, as they are instance triples.

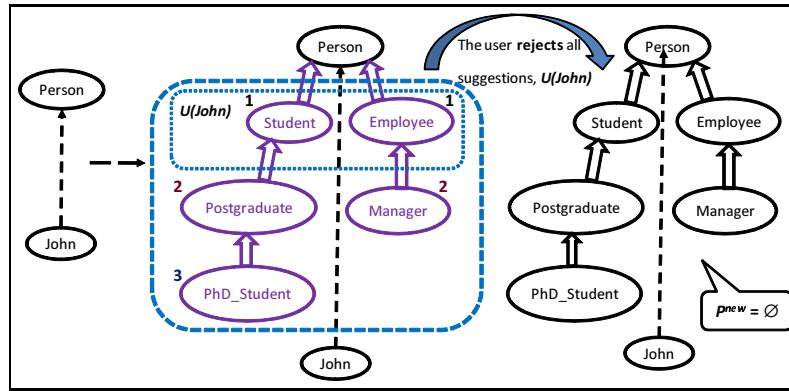


Figure 4: One migration and rejection of the computed recommendations

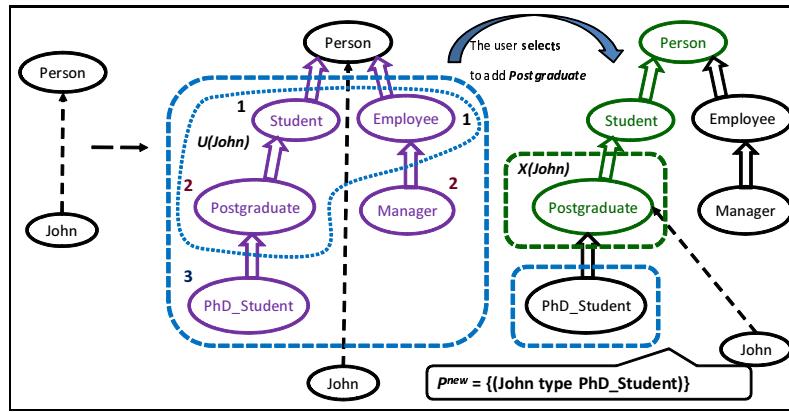


Figure 5: One migration and acceptance of some recommendations

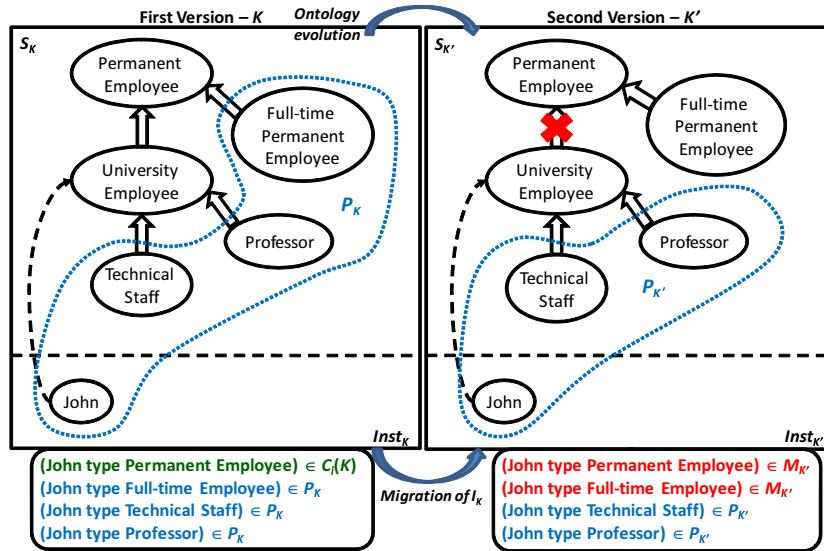


Figure 6: Migration to a non backwards compatible schema

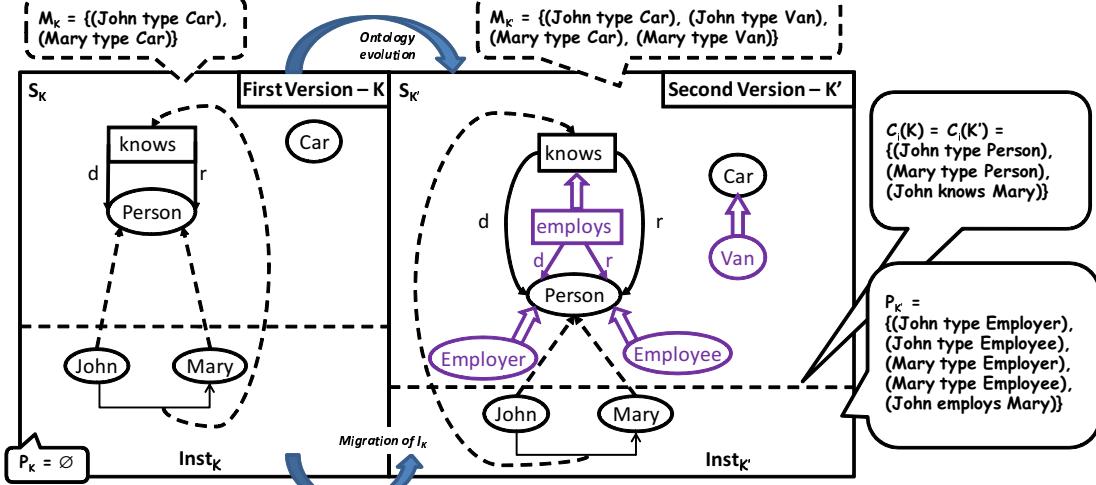


Figure 7: Examples with properties

deleted (explicit and inferred) schema triples in the destination ontology. The user can also get information about the possibilities of the source and the destination ontology, e.g. the number of possible class instance triples and possible property instance triples in both ontologies.

To curate the resulting descriptions (“Curate” menu), RIMQA allows the user to inspect all possible class and property instance triples. Regarding class instance triples, all possible class instance triples are listed and the user is able to add (by pressing the “Accept” button) one or more possible class instance triples to the certain part of the extended KB (eKB). Subsequently, the selected possible class instance triples and all their supertriples are added to the certain part of the eKB and they are removed from the multiple choice list and from the possible part of the eKB . The user can also remove (by pressing the “Reject” button) one or more possible class instance triples from the possible part of the eKB . Subsequently, the selected possible class instance triples and all their subtriples are removed from the multiple choice list and from the possible part of the eKB . After that, the user selects to save the new certain and possible part of the eKB (by pressing the “Save eKB ” button).

If the user selects to save the eKB (by pressing the “Save eKB ” button), we store the new instance triples, i.e. the certain part of the eKB , in a .rdf file (called “newCertainModel.rdf”) and the new possible instance triples, i.e. the possible part of the eKB , in a different .rdf file (“newPossibleModel.rdf”).

4. CONCLUDING REMARKS

The rapid evolution of ontologies requires principles, techniques, and tools for managing the quality of the migrated descriptions, as well as flexible interactive methods for managing the incurred uncertainty. To the best of our knowledge this is the first work that exploits ontology schema evolution for managing the specificity of instance descriptions. According to our opinion this is key issue for the preservation of scientific data, i.e. for e-Science.

Since the ultimate objective is not just the identification

of possibilities, but to aid making the instance descriptions as specific as possible, we proposed a specificity *lifecycle management process* that *ranks* the possible instance triples, prompts to the user a subset of the possible instance triples and we show how the extended KB should be *updated* when the user *accepts* or *rejects* some of them. To investigate the feasibility of our approach, we designed and developed a prototype system.

There are several issues for future research. One interesting direction is to generalize our approach to the *XSD⁵*-typed literal values [12] of property instance triples. Such extension would allow reasoning about the *accuracy* of the migrated descriptions over linearly ordered domains (e.g. as consequence of migrating 32-bit floating numbers to a 64-bit representation).

Acknowledgements

Work done in the context of NoE APARSEN (Alliance Permanent Access to the Records of Science in Europe, FP7, Proj. No 269977, 2011-2014).

5. REFERENCES

- [1] C. E. Alchourrón, P. Gärdenfors, and D. Makinson. On the Logic of Theory Change: Partial Meet Contraction and Revision Functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.
- [2] C. Bizer, T. Heath, and T. Berners-Lee. Linked data—the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.
- [3] D. Brickley and R. V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation, February 2004. Available at <http://www.w3.org/TR/rdf-schema/>.
- [4] Peter Buneman, James Cheney, Wang Chiew Tan, and Stijn Vansumeren. Curated Databases. In *27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS-2008)*, pages 1–12, 2008.

⁵XML Schema Definition

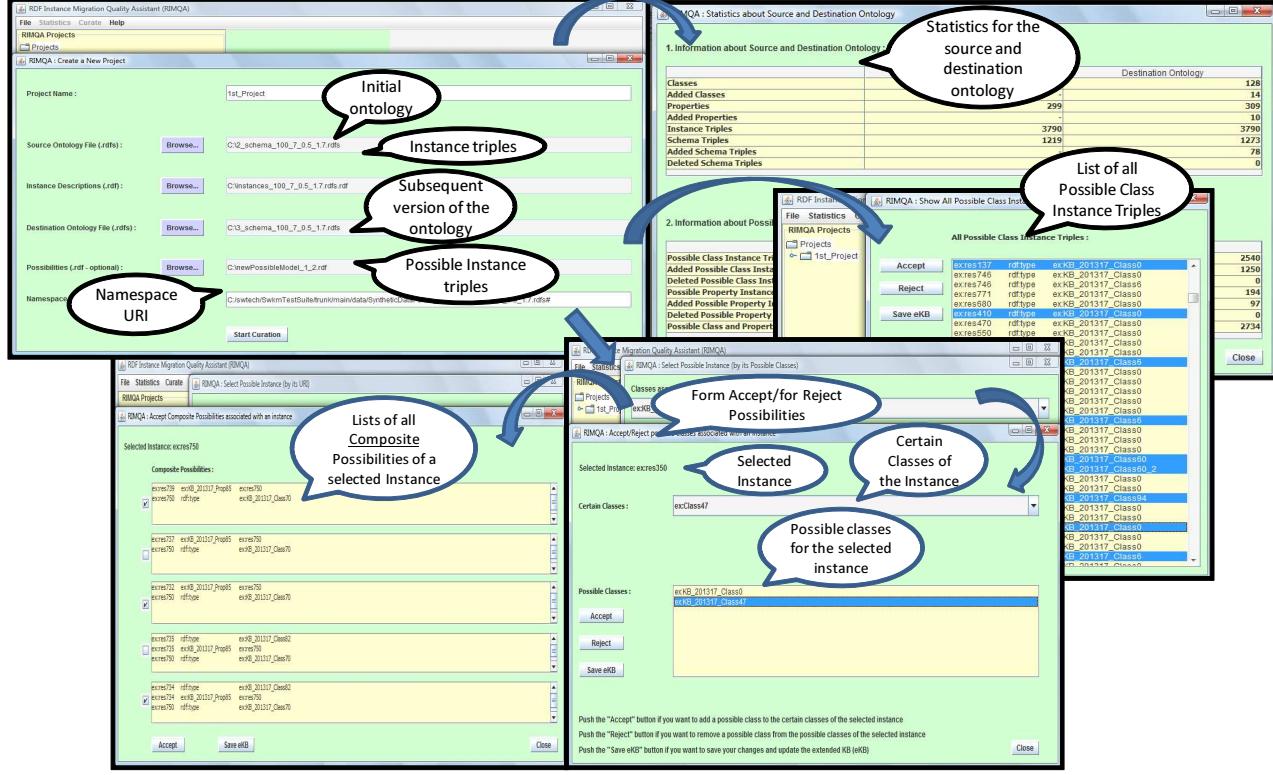


Figure 8: Some screenshots of the GUI of RIMQA

- [5] M. Dalal. Investigations into a Theory of Knowledge Base Revision: Preliminary Report. In *7th National Conference on Artificial Intelligence (AAAI-1988)*, pages 475–479, 1988.
- [6] M. Doerr and Y. Tzitzikas. Information carriers and identification of information objects: An ontological approach. *Arxiv preprint arXiv:1201.0385*, 2012.
- [7] G. Konstantinidis, G. Flouris, G. Antoniou, and V. Christophides. A Formal Approach for RDF/S Ontology Evolution. In *18th European Conference on Artificial Intelligence (ECAI-2008)*, pages 70–74, Patras, Greece, July 2008.
- [8] Y. Marketakis, M. Tzanakis, and Y. Tzitzikas. PreScan: Towards Automating the Preservation of Digital Objects. In *Procs of the Intern. Conf. on Management of Emergent Digital Ecosystems MEDES'2009*, Lyon, France, October, 2009.
- [9] Y. Marketakis and Y. Tzitzikas. Dependency Management for Digital Preservation using Semantic Web technologies. *International Journal on Digital Libraries*, 10(4), 2009.
- [10] Thomas Neumann and Gerhard Weikum. x-RDF-3X: Fast Querying, High Update Rates, and Consistency for RDF Databases. *Proceedings of the VLDB Endowment (PVLDB)*, 3(1):256–263, 2010.
- [11] Natalya Friedman Noy and Michel C. A. Klein. Ontology Evolution: Not the Same as Schema Evolution. *Knowledge and Information Systems*, 6(4):428–440, 2004.
- [12] David Peterson, Shudi (Sandy) Gao, Ashok Malhotra, C. M. Sperberg-McQueen, and Henry S. Thompson. W3C XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes, W3C Working Draft 3, December 2009. Available at <http://www.w3.org/TR/xmlschema11-2/>.
- [13] Li Qin and Vijayalakshmi Atluri. Evaluating the validity of data instances against ontology evolution over the Semantic Web. *Information & Software Technology*, 51(1):83–97, 2009.
- [14] Satya S. Sahoo, Wolfgang Halb, Sebastian Hellmann, Kingsley Idehen, Ted Thibodeau Jr, Soren Auer, Juan Sequeda, and Ahmed Ezzat. A Survey of Current Approaches for Mapping of Relational Databases to RDF, 2009. Report by the W3C RDB2RDF Incubator Group. Available at http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf.
- [15] M. Theodoridou, Y. Tzitzikas, M. Doerr, Y. Marketakis, and V. Melessanakis. Modeling and Querying Provenance by Extending CIDOC CRM. *J. Distributed and Parallel Databases (Special Issue: Provenance in Scientific Databases)*, 2010.

Package Formats for Preserved Digital Material

Eld Zierau

The Royal Library of Denmark
Søren Kierkegaards Plads 1
1219 København K
ph. +45 33 47 46 90

elzi@kb.dk

ABSTRACT

This paper presents an investigation of the best suitable package formats for long term digital preservation. The choice of a package format for preservation is crucial for future access, thus a thorough analysis of choice is important.

The investigation presented here covers setting up requirements for package formats used for long term preserved digital material, and using these requirements as the basis for analysing a range of package formats.

The result of the concrete investigation is that the WARC format is the package format best suited for the listed requirements. Fulfilling the listed requirements will ensure mitigating a number of risks of information loss. Thus WARC is the best choice for a package format in cases where these same risks are judged most important. Similar analysis will need to be carried out in cases where the requirements differ from the ones described here, e.g. if there are specific forensic or direct access to files.

Categories and Subject Descriptors

E.2 Data Storage Representations: Linked representations, Object representation

E.5 Files: Backup/recovery, Optimization, Organization/structure

H.3.7 Digital Libraries: Collection, Standards, Systems issues

I.7.1 Document and Text Editing: Document management, Version control

I.7.2 Document Preparation: *Format and notation, Standards*

General Terms

Management, Documentation, Design, Standardisation.

Keywords

Package formats, Digital Preservation, Bit preservation.

1. INTRODUCTION

This paper presents an investigation of different possible package

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

iPRES'12, October 1–5, 2012, Toronto Canada.

Copyright 2010 ACM 1-58113-000-0/00/0010...\$10.00.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. iPRES'12, Oct 1-5, 2011, Toronto, ON, Canada. Copyright 2012, Digital Curation Institute, iSchool, University of Toronto.

formats that can be used for packaging digital material for long term preservation. The investigation has resulted in suggesting the WARC format as the package format to be used for bit preserved digital material at The Royal Library of Denmark [2].

The selection of a package format for digital material is crucial for how to facilitate long-term accessibility. The selected package format is used to package files that must be sent to bit preservation, which must ensure that the bit-streams remain intact and readable [11,25]. That means the package format will constitute the frame of the digital material, and thus be the basis for general recovery of data and future data access as well as functional preservation actions of the original bits, where functional preservation ensures that the bits remain understandable and usable according to the purpose of preservation [25]. A package formats is presumed needed, because files must be applied a minimum of metadata in terms of an identifier as described later.

The topic of long term preservation package formats has partly been treated in a recent paper: “Digital forensics formats: seeking a digital preservation storage format for web archiving” [10]. As the paper states: “There has been little consensus on best practices for selecting storage container”. The paper presents an overview of archiving formats for digital forensics that can satisfy the requirement of tracing originality. This present paper on the other hand will not focus on requirements for forensics, but instead will focus on requirements for long term preservation in general.

The goal of the investigation was to find as few suitable package formats for packaging for as many types of different materials as possible. The reason for this goal is that each package format will require resources in form of skills and documentation in order to maintain accessibility to the material. Thus in order to minimize costs and in order to minimize the risk of losing skills for a specific format, the number of formats must be kept as low as possible.

Diverse types of digital material can for instance be found for libraries. Libraries usually have many types of different digital materials that are candidates for long term preservation. For instance substitution copies of analogue materials [9]; harvested web material [2]; emails from authors and forensic images of e.g. author's hard discs [10]. The digital material can consist of different files with different file formats and metadata, and the material can be composed digital objects (called representations as in PREMIS terminology [17]) with various metadata.

This paper will argue for a set of requirements that should be considered in choice of a package format used in long term preservation of diverse types of digital material. Such

requirements will depend on the purpose of the preservation, the nature of the material to be preserved and individual prioritization of risk that must be mitigated by the way the material is preserved. Thus, the given requirements are arguable requirements to be considered, while the weight of meeting them can differ.

The next section will provide the general requirements for a package format used for long term preservation of digital material. The following section ‘Alternative package format choices’ describes a range of packaging formats and analyses how they meet the different requirements for a package format.

2. FORMAT REQUIREMENTS

The format requirements described here are the requirements for formats used for archive packages under long term preservation. The following contains descriptions and argumentations for a number of such requirements. These requirements are either related to the actual packaging and storage, to preservation aspects, or to identification of contents of packages.

2.1 Package and storage related requirements

The following requirements are requirements related to packaging and storing. These are selected requirements which cover the most often referred requirement about independence, as well as requirements related to flexibility concerning exploitation of storage resources. More detailed requirements are left out in order to give a comprehensive presentation (additional requirements can e.g. be found in [2,10]).

Requirement 1: Independence of storage platform

For long term bit preservation, data will in most cases be stored on different media using different operating systems. This is, for instance, the case for one material in order to ensure independence between copies of data in a bit repository, which takes care of holding and preserving bits [25]. In the long term this is likely to be the case at some stage as a consequence of changes in storage technology. Thus a basic requirement for a package format used in long term bit preservation is: *The Package format is independent of storage platform* [2], which has been formulated in many ways as a requirement for sustainable file formats in general [2,10,12,13,14,22].

Requirement 2: Package format allows flexible packaging

A requirement related to how well the format can support optimization of storage use is: *Package format allows flexible packaging*. This can relate to economical or performance related issues concerning the best way to package, making different sizes of packages. There can be benefits in having large packages according to how the storage works. On the other hand there can be accessibility issues which can mean that smaller packages are preferred. Reasons to keep to small packages can be technology changes as well as challenges in having different parts of the packages with e.g. different confidentiality levels. Anyhow, flexibility will mean that package sizes can be optimized according to chosen policies¹.

¹ Discussion on this subject is documented in mail correspondence with Kevin Ashley on the JISC Digital Preservation mailing list. Please refer to <https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1105&L=digital-preservation&F=&S=&P=7686>

Requirement 3: Allow update records

A requirement related to the ability to minimize needed storage volume is to require that the: *Package format allows update records*. Since data packages for long term preservation are static, they cannot be changed after bit preservation has started. Therefore the only alternative to update packages is to make a full new representation and bit preserve this representation as well. However, in many cases this can be expensive, for instance in the case where a large TIF file has a single letter change in the TIF file header. However, the opportunity of having update records must be carefully considered in terms of the complexity it can add to the long term interpretation of the data.

2.2 Preservation related requirements

Preservation related requirements for package formats cover aspects of ensuring that the packages are readable and understandable in the future. These have many similarities to general requirements for preservation file formats [12,13,14,18,22]. Common to such requirements is that they are related to mitigating risk such as losing information in the digital material or losing ability to interpret the information [20,24].

The following requirements are deduced from an analysis where risks and requirements are considered for digital material that will have a large variety and will have to be long term preserved. These requirements are based on the above mentioned literature and further details can be found there as well.

Requirement 4: Must be Standardised format

The first requirement is that it: *Must be a standardised format*. This covers the degree to which the format has gone through a rigorous formal standardisation process [12,13,14,18,22]. This relates to the future ability of thorough and accepted documentation for the format which will mitigate risk of losing means to understand the format.

Requirement 5: Must be open

A related requirement is that a format: *Must be open* [2,14,18,22]. This requirement relates to risks of losing the ability of future interpretation of the format. If the format is not open, there may arise legal and economical issues concerning tools to interpret the contents of the format. Furthermore, there may be a risk that documentation of the format is unavailable after e.g. copyrights of the format have expired.

Requirement 6: Must be easy to understand

Another related requirement is that the format: *Must be easy to understand*. This requirement is usually referred to in connection with transparency [2,12,13] and complexity [6]. The requirement relates to the future ability to understand the package format, and to mitigate the risk of introducing errors or later difficulties in interpreting the contents of packages. This risk is high if the format is too complicated.

Requirement 7: Must be widely used in bit repositories

There is a requirement stating that the format: *Must be widely used in bit repositories*. This covers ubiquity in terms of the extent to which the format has been adopted. In particular in this paper *widely used in bit repositories* means the extent of adoption by national libraries, archives, and other memory institutions internationally [12,13,14,18,22].

Requirement 8: Must be supported by existing tools

A related requirement is that the format: *Must be supported by existing tools*. This also concerns the trust in quality and future existence of the format, which again will mitigate the risk of losing ways to understand the format in the future. Furthermore it concerns the ubiquity aspect in terms of how widespread the format can become [14,18].

Requirement 9: Must be able to include digital files unchanged

The final preservation format related requirement is that the format: *Must able to include digital files unchanged*. This requirement addresses mitigation of the risk of losing information as a result of changes made to files in the packaging process. Such changes could for instance occur in connection with compression (partly discussed in [12,22]). Or in cases where the package format is XML based, and conversions are needed in order to include files in XML structures due to the fact that XML is tag based, and end tag can be part of the files.

2.3 Identification related requirements

The last requirement covered in this paper is a requirement related to the ability to identify contents of packages, which is the basic metadata of any digital piece of information.

Requirement 10: Must facilitate identifiers for digital files

The requirement that a package format: *Must facilitate identifiers for digital files*. This requirement is related to more general requirement of flexibility of embedding metadata [10]. It does however deserve special attention and explanation, since it is crucial for future reference of files which are part of digital material.

In general we have three different types of data which must be recorded in packages. The three different types of data² are:

- *Digital files* of any file format will need to be addressed in different contexts, such as metadata for the file or relations to the files as part of a digital object. Therefore the digital files must be identifiable. This is done by assigning an identifier to each file.
- *Metadata to digital files* as metadata about the files separated from the actual files. This metadata will as a minimum consist of the identifiers for the digital files.
- *Metadata for a representation*. All information for contexts and metadata can be put into e.g. a METS³ structure with references to the involved files and metadata.

These types correspond to the object types ‘file’ and ‘representation’ in the PREMIS metadata standard, where a representation can be purely representation of file metadata.

Different metadata schemes facilitate definition of identifiers for the metadata, thus it is no problem to make schemes of how to represent identifiers for and within the metadata. However, definition and attachment of usable identifiers for digital files is a challenge, since the digital file itself may not carry the information of the identifier of the file.

² Except from the metadata part, this corresponds to different types of PREMIS objects [16]

³ Metadata Encoding & Transmission Standard (METS) <http://www.loc.gov/standards/mets/>

One solution to meet this challenge could be to simply place the files as bit chunks with the identifier to the bit repository, and leave it to the bit repository to make the connection between the file and the identifier. However the information that the file has been assigned the specific identifier is also crucial for long term preservation. If we leave it 100% up to the bit preservation solution to preserve the link between files and identifiers, we will risk that we cannot recreate the data in case this index is lost. Furthermore, if the identifier is only expressed as an identifier in a bit repository, we eliminate any optimisation of packaging more files or files and metadata in the same packages for a bit repository. Therefore the best way to ensure the relation is to put the identifier with the file.

There are different ways to assign information of an identifier with a file:

- *Naming files with the identifier*

Using identifiers in file names is generally not considered a good solution, for a number of reasons:

Firstly, because there can be restrictions to how files are named which can conflict with the general scheme to name persistent identifiers.

Secondly, because a file name is not part of the file itself, it is information of the file system. Furthermore, the file name can only be unique in connection with a file path anyway, and a file path will include an assumption on how files are placed which is likely to change in a time frame of 50 years. This again can give challenges to update of reference and resolver schemes.

Thirdly, file names may not make sense in the future, and in a bit preservation context with different copies on different media as e.g. microfilms, file names may not exist or may be different for different copies in a bit preservation system.

- *Put identifier into files as inherited metadata*

Insertion of an identifier into files would have to be done before the files are sent to bit preservation. This could work for some cases, but cannot be used in all cases. First of all because not all file formats allow inherited metadata. Secondly, because there may be requirements to leave the file untouched (e.g. a forensic disc image). In general it would also require knowledge of how to extract the identifier from all bit preserved file formats, which in practice would not be possible for collections with all types of digital material.

- *Wrap files and identifier in a package format*

Wrapping an identifier with the file in a package will set requirements for the abilities of the package format, since this is not a trivial feature that applies for all package formats.

This requirement of facilitating identifiers for digital files is therefore based on the assumption that we want to mitigate the risk of losing identifier information because of environment or file format dependencies.

3. FORMAT CHOICES

This section describes a range of different package formats that could be candidates for a general package format for a wide range of digital material, as is usually the case for libraries. This section will furthermore describe how well the formats fulfil the different requirements listed in the previous section.

3.1 Considered package formats

The following considered package formats are chosen based on knowledge of package formats used in other libraries and archives repositories⁴, formats described in the paper “Digital forensics formats: seeking a digital preservation storage format for web archiving” [10], and generally known package formats such as ZIP and RAR. The list of formats does not constitute an exhaustive list of formats. For instance the Archive eXchange Format (AXF)⁵ is excluded since “... it is a very new development, with a lack of access to detailed documentation and source code, making it difficult to assess” [10]. Also formats for very specific purposes like the optical media disk imaging format iso image are excluded [8], and the format gzip⁶ which is a compression format and thus cannot fulfil the requirement of unchanged files. In order to narrow the list, there are also formats that are described together with other formats, which for instance is the case for XFDU which is mentioned under METS.

3.1.1 AFF

Advanced Forensic File Forensic disk image formats such as AFF⁷ and AFF4⁸ are formats specifically designed for to contain metadata for forensics. These formats have the benefit of providing settings to control the quality, speed, and size of output data. One disadvantage of AFF is that it assumes that the image is from a disk as opposed to a collection of files or folders [10].

Take for example the AFF4 format, an open format which is proposed to be adopted as a standard evidence management platform [3]. The AFF4 is a position based format with the ability to insert specific forensic metadata. However it does not support means of update records.

3.1.2 ARC

The ARC format is a position based format originally designed for web archiving packages. It is based on record definitions identified by name tags and byte length. It requires that the first record in a package is a header record, a ‘filedesc’ record, with information that is only used in the context of web archives and thus can add confusion and take up space for packages that are not web archive specific⁹ [11].

The ARC format has a fixed set of record definitions, i.e. it does not include the possibility to define separate update records. The ARC format is not described in a standard and it is not very widely used for other archives than web archives. Furthermore, there is a tendency that web archives using ARC are moving to use WARC instead [23].

⁴ Partly based on the previously mentioned mail on the JISC Digital Preservation mailing list

⁵ See <http://www.openaxf.org/> for description of AXF

⁶ The gzip fomat is defined in “GZIP file format specification version 4.3”, <http://www.ietf.org/rfc/rfc1952.txt>

⁷ See description of Advanced Forensics Format (AFF) on <http://www.forensicswiki.org/wiki/AFF>

⁸ See description of Advanced Forensics Framework 4 (AFF4) on <http://www.forensicswiki.org/wiki/AFF4>

⁹ See “Arc File Format, Version 1.0”, <http://www.archive.org/web/researcher/ArcFileFormat.php>

3.1.3 BagIt

The BagIt¹⁰ format is intended for quick packing and unpacking into folders. It was originally design for exchange of information, i.e. BagIt is not directly designed for packaging to archives. The BagIt format only provides a way to specify certain metadata to a package, whereas the package itself must be specified to be a package in e.g. TAR or ZIP formats.

The BagIt format provides a structure for how files can be packed in e.g. a TAR or a ZIP file. It allows for specification of one external identifier, but otherwise it does not offer other ways to address the files in the bag aside from their file names.

The BagIt format is used both as exchange format but also as a package format for data in a repository¹¹. The BagIt format is not formally standardised. The BagIt format cannot be extended with support of update records.

3.1.4 METS

The Metadata Encoding and Transmission Standard (METS) specifies an XML based format which originally was designed for transmission of information, but is today widely used as a container format for metadata to digital material¹² [22].

The METS format could in theory be used as a package format, although there are challenges regarding inclusion of digital files in a METS structure. The challenge is due to the fact that METS is an XML based format and in practice XML is not suited for inclusion of digital files, since objects are defined via start and ending tags. Thus the file will need to be transformed in order to avoid ambiguity in case the file itself includes bit sequences that can be interpreted as an end tag. This is probably the reason why METS is often used as metadata format but rarely used as the actual package format (examples of METS packed in WARC or BagIt can be found in [5] and [4]).

The METS format is very flexible and can include a range of other XML based metadata formats. It may therefore be possible to exploit this flexibility to include specification of update records. The METS format is a widely used standard hosted at the Library of Congress¹³. However, the standardisation is related to METS as a metadata standard rather than a package format standard.

Another similar format is the XFDU format [1], also an XML based metadata format. The XFDU format therefore has the same challenges as METS also being based on XML.

3.1.5 RAR

RAR stands for Roshal ARchive. It is a proprietary archive file format that includes data compression¹⁴. The RAR format is not an open format and it is not formally standardised.

¹⁰ The BagIt fomat is defined in “The BagIt File Packaging Format (V0.97)”, <http://tools.ietf.org/html/draft-kunze-bagit-06>

¹¹ See e.g. <http://www.dcc.ac.uk/resources/external/bagit-library>

¹² See e.g. “METS Implementation Registry”, <http://www.loc.gov/standards/mets/mets-registry.html>

¹³ See <http://www.loc.gov/standards/mets/>

¹⁴ See “RARLAB” for description of the RAR format <http://www.rarlab.com/>

RAR files may be created only with the commercial software WinRAR, RAR, and software that has been granted permission.

The RAR format is mainly focused on technical issues related to the actual storage of packages in compressed form. It does not provide means to specify external identifiers and there are no possibilities of making extensions with update records.

3.1.6 TAR

The TAR format¹⁵ provides a way to package file folders and their contents. The TAR format is file oriented, but also byte oriented. The TAR format has no centralized location for the information about the contents of the file, i.e. it is not easy to make relations between identifiers and files. The best way to assign identifiers to TAR elements is to use the BagIt format which opens more possibilities to specification of different data.

The TAR format is a standardised (POSIX.1-2001) format which is widely used for archiving of tapes in general, and there are different tools available for the format. The TAR format does not support the notion of update records.

3.1.7 WARC

The WARC format is a position based format focused on web archiving, but has a general design which can also be used for other purposes, leaving out web specific information [7].

The WARC format consists of different record types, where a record e.g. can contain a file as well as record information as for instance the identifier for the record/file. Thus WARC provides an easy way to assign an identifier to a file.

The WARC format has recently been ISO standardised [7], but is not used very widely yet and there are few tools available. WARC has recently been used for other material than web material in the German Kopal project [21].

As for the ARC format, the WARC format also has header information, but in this case it can consist of information that is relevant for a bit repository, including the identifier for the package itself.

There have been initiatives to develop tools for WARC in different contexts: at the University of Maryland¹⁶, in an IIPC project¹⁷, and at Internet Archive¹⁸. However, these tools are still not mature enough to consider as proper production tools [15].

The standard includes the possibility to define your own record type [7], which enables us to specify updates as basis for update mechanisms.

¹⁵ Description of the tar file format can be found on <http://www.gnu.org/software/tar/>

¹⁶ See “An Approach to Digital Archiving and Preservation Technology – WarcManager”, <https://wiki.umiacs.umd.edu/adapt/index.php/WarcManager>

¹⁷ See “Open Source WARC Tools - Functional Requirements Specification”, http://warc-tools.googlecode.com/files/warc_tools_frs.pdf

¹⁸ See “Release Notes - Heritrix 3.1.0-RC1”, <https://webarchive.jira.com/wiki/display/Heritrix/Release+Notes+-+Heritrix+3.1.0-RC1>, retrieved October 2011

3.1.8 ZIP

The ZIP file format¹⁹ is a file format, which is used for data compression and as an archive format, which also allows for uncompressed packaging. A ZIP file can contain file folders and files. For each entry there are defined a number of fields like file name, compression algorithm etc. The format also allows specification of additional fields, e.g. the identifier for a file.

The ZIP format was originally published as an open format [16]. Although ZIP is widely used in general and proposed to be standardised, it has never been formally standardised²⁰. Furthermore it should be noted that although ZIP is widely used in general, it is not as common to see ZIP used as package format in archives and libraries.

There are different implementations and interpretations of the ZIP format [10]. Exploiting the ability to define an identifier in an extra field would also require specifically design zip tools to make this information extractable.

The ZIP format does not have any direct mechanism enabling introduction of update records.

There are different software components deployment formats building on ZIP, e.g. the Web application ARchive (WAR)²¹ file format, and the Java Archive (JAR)²² file format. As these formats are designed for software deployment rather than for archiving, these formats do not provide extra means for archiving than the ZIP format.

3.2 How the formats meet requirements

An overview of how the presented package formats meet the requirements for the package format used in long term preservation is provided in table 1. The table provides approximate ranking of how well the formats meet the requirements. These rankings are expressed by the five ranking values (illustrated by colours in order to give a better overview):

<i>Yes</i>	if the requirement is considered to be sufficiently met
<i>Almost</i>	if the requirement almost can be considered to be sufficiently met, but not completely
<i>So-So</i>	if the requirement is considered to be met to some extent, but thorough evaluation of deficiencies is required
<i>Little</i>	if the requirement is only considered to be sufficiently met to a minor degree
<i>No</i>	if the requirement is not considered to be met at all

The ranking is only approximate values, since e.g. definition and evidence of whether formats are widely used are only based on

¹⁹ See “ZIP File Format Specification” <http://www.pkware.com/documents/casestudies/appnote.txt>

²⁰ See <http://www.itsci.ipsj.or.jp/sc34/open/1414.pdf> which proposes standardisation.

²¹ See e.g. “Web Application Archives” for description of the Web ARchive (WAR) file format (Sun), http://java.sun.com/j2ee/tutorial/1_3-fcs/doc/WCC3.html

²² See e.g. “JAR File Specification” <http://docs.oracle.com/javase/6/docs/technotes/guides/jar/jar.html>

knowledge of a small set of larger institutions. It should also be noted that there is an emphasis of use of the formats as package formats in preservation, thus the METS format is rated to be ‘so-so’ *widely used in bit repositories*, since it is widely used as a metadata format, but not as a package format. Likewise the ZIP is ranked ‘so-so’, since the requirement concerns the widespread use

of ZIP with bit repositories for long term preservation in larger preservation institutions. Another example of approximation is that the BagIt format cannot offer flexible packaging when the external identifier for a bag is used as identifier for a file, since this means that a bag can only include one file.

Table 1. Package formats fulfilment of requirements

Requirements \ Formats	AFF	ARC	BagIt	METS	RAR	TAR	WARC	ZIP
1. Platform independent	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
2. Flexible packaging	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
3. Supports update packages	No	No	No	Almost	No	No	Yes	No
4. Standardised	Little	No	So-so	Yes	No	Yes	Yes	Little
5. Open	Yes	Yes	Yes	Yes	No	Yes	Yes	Almost
6. Easily understandable	So-so	So-so	So-so	Almost	No	Little	Yes	Little
7. Widely used in bit repositories	No	So-so	Almost	So-so	Little	Yes	Almost	So-so
8. Tools available	So-so	Yes	Yes	So-so	Yes	Yes	So-so	Yes
9. Include files unchanged	Yes	Yes	Yes	No	No	Yes	Yes	Yes
10. Identifiers for files	Yes	So-so	So-so	Yes	No	No	Yes	No

3.3 Suggested choice of WARC

The requirements ranked in table 1 should not be equally weighted. First of all the importance of long term preservation is regarded as highest. Secondly, there are requirements that become less important, if other requirements are given high score. For instance, it may not be important that a format is *Standardised*, in case the format has high scores on *Easily understandable*, *Open* and *Widely used*. Such a format may have a higher chance of surviving as a de facto standard, than another standardised format which is neither *Easily understandable* nor *Widely used*. Similar for tooling, a format that is *Open* and *Widely used* is quite likely to get *Tools available* in a relatively short time.

The final suggestion of WARC is therefore based on analysis that takes such considerations into account, and using exclusion of formats by comparison between the formats.

ARC can be ruled out, since it is a much more primitive and immature package format than WARC, thus arguments for choosing ARC will also be arguments for choosing WARC, but WARC has more benefits than ARC.

METS and **XFDU** can be ruled out, since they are XML based which cannot support proper inclusion of files, which is crucial and thus a mandatory requirement for the long term preservation.

RAR is ruled out since it can only offer compressed packaging which cannot be accepted for *all* long term preservation.

If the requirement to assign identifiers for files is considered crucial, then the **TAR** and **ZIP** formats are best considered in connection with the BagIt format. From table 1 it is evident that the TAR format better fulfils other requirements, since it has the same score or better score than the ZIP format for the same requirements.

The only real problem with **BagIt** is that it only can have one external identifier assigned to a package, which is probably due to the fact that it is designed as an exchange format. This fact means

that settling for BagIt would limit the possibilities of how to make packages, since use of external identifiers for identifiers means that a bag can only have one file. However, it only has low ranking of requirements that are considered less important for long term preservation, and it is therefore worthwhile to consider this format. However, besides BagIt, there will have to be a decision on whether it should build on TAR or ZIP.

The **WARC** format is a candidate since it can support all requirements, although it is not widely used yet (at least as package format for all types of digital material), and there is no stable tool package to support it. However, there are a lot of indications that this will change to the better, since web archives will start to use WARC instead of ARC. Furthermore, using WARC for other than web material is not entirely new. For instance the German Kopal project is today working towards packaging all types of materials in WARC when sent to bit preservation [21] (using Private LOCKSS Networks [19]).

Finally the AFF format could be a candidate, but compared to BagIt and WARC, it loses on the fact that there is limited experience in use as a general package format, and is not widely used. As presented in the [10] WARC only lacks the ability to represent file system structure or the file system characteristics in order to meet requirements for forensic data. However, in the preservation perspective taken in this paper, this is not crucial, since such metadata can just as well be part of the packed metadata.

The two most relevant alternatives found in the analysis are therefore WARC and BagIt based on TAR.

The only requirement where WARC scores lower than BagIt is the same requirement as the lowest score for WARC, namely: *Tools available*. This means that there may be a risk that local investments must be made for tools using WARC. However, the interest in using WARC for web archiving indicates that a community for tool development exists and tools probably will emerge soon.

The two formats have the same score *Widely used*, but for different reasons. Although BagIt is designed as an exchange format, it is also used for repository material. WARC on the other hand is mostly used for web archive material, or is most likely to be used in most future web archives. The risk that they may not go for the WARC format after all is quite slim, since WARC is now both the only formally standardised format for web archiving, but also the best alternative, since it is developed based on previous experiences with web archiving formats like ARC.

Great advantages with WARC compared to BagIt are that it can represent *Identifiers for files* easily, and a WARC package is in easily understandable text form. On the other hand BagIt can only represent one external identifier per bag and interpretation relies on knowledge of both BagIt and TAR.

The restraints on how to use external identifiers in BagIt also mean that the WARC format is best with regard to flexible packaging. This enables the possibility of choosing to put metadata for files in the same package as the file, or even more objects in the same package. As the size of packages can have impact on different resource issues the flexibility in settling for policies in using WARC can affect optimization resource use.

Finally the WARC format is the only format of the mentioned ones²³, where it is possible to define update records directly. This is not the most crucial requirement, but it can help to optimise preservation costs, if the risk analysis from bit preservation can allow preservation of updates as an alternative way of preserving a representation.

Besides the advantages that WARC have considering the requirements, WARC also has an extra advantage for institutions with web archives using WARC: The institution will only need skills concerning WARC as package format for all preserved data. This is for instance the case for The Royal Library of Denmark. It should however be noted that the way WARC is used for web archives may be more advanced than the way WARC is used for other materials. Still it is a great advantage not to need skills for more package formats.

A discussable advantage of WARC is that it does not rely on assumptions of having folder and file structures. As expressed in "Cedars Guide to: Digital Preservation Strategies"²⁴.

"The UNIX format known as tar (originally standing for tape archive) is used by Cedars as the preservation byte-stream for such cases, because it is publicly documented, and there exists public domain software for writing and reading data in such a format. Another institution may choose to use a different format for mapping the original file tree into a byte stream. Whatever format is chosen, it must enable a subsequent recreation of a file system that operates in the same way as the original. Thus the files system should be converted to a byte-stream for preservation by use of tar or other suitable program."

In other words TAR does have assumption of file and folder structures as the basis for unpacking the TAR file. Whether this

²³ Other formats supports update specification, e.g. VCDIFF (<http://tools.ietf.org/html/rfc3284>), but these are typically not suited as general package formats

²⁴ See <http://www.imaginar.org/dppd/DPPD/146%20pp%20Digital%20Preservation%20Strategies.pdf>

will exist in 100 years can only be a guess, thus there will be different opinions on whether risk of losing the basis for unpacking TAR files should be included in a risk analysis as basis for choosing a package format.

4. DISCUSSION

It could be argued that this paper should have included a more complete list of formats that can be used for packaging data that are to be bit preserved. However, most other alternatives are less known formats, commercial formats or formats designed for a specific purpose. Thus such formats would most likely be eliminated on requirements of being open, standardised and widely used.

This paper has only included the most relevant requirements for preservation of general digital materials. There can be supplementary requirements for e.g. how the format supports availability of data. Such requirements are described in the literature consisting of guidelines, reports and papers [2,10,12,13,14,18,22].

The requirement of expressing *Identifiers for files* is crucial for the choice of WARC in the present presentation. Therefore there may be cases where such an analysis will not lead to the same result. This would for instance be the case where this requirement is seen as less important, due to e.g. relying on a bit repository to keep track of the identifier, having few formats where risk of losing embedded identifiers is seen as unimportant, or risk related to having identifiers as part of the file name is considered minor.

Another example, where analysis of choice for package format is different, is the package format for forensic digital material as given in [10]. This is due to the fact that the requirements and focus are different. It may be that the choice of package format will be different for different types of digital material, e.g. forensic and other digital material. However, it should be noted that there are no limitations in WARC to include AFF packages. This could be desirable in the case of the benefits of a general package format in a bit repository, e.g. in order to have similar access to all packages. However it can also be considered more beneficial to have several package formats, since overhead in unpacking, and possibly impact of access time of the data can be avoided. Likewise, there could be other specific digital material that needed specific considerations, e.g. specific scientific data.

The packaging for bit preservation may not be optimal for the way digital material is e.g. disseminated. The focus is on preservation. Thus the focus regarding availability is that it will be possible to reproduce digital material and identifiers, solely based on the preserved packages. This means that additional analysis will be required for cases where there are specific requirements to access time that are more important than preservation requirements.

5. CONCLUSION

This paper found the best suited format for long term preservation of varied digital materials is WARC. However, the value of the analysis depends on whether the presented requirements are seen as the most important requirements for the digital preservation of the material, and whether there are other requirements to be included.

Compared to most other formats, the WARC format is strong as a preservation packaging format in general, especially regarding issues of: applying identifiers to bit-sequences/files, being easily understandable and being one of the few formally standardised

formats. Furthermore the WARC format is the only format among the listed formats that is extendible with record definition for update records, which can give economical benefits for preserving changing materials.

The only point where the WARC format does not have the top score is how widely used the format is, and how well it is supported by tools. However, the lower score concerning ‘widely used’ is based on the fact that it is mostly used within web archiving, although there are no restrictions or overhead in using the WARC format for other types of digital archiving. Regarding tool support, the increasing use of the WARC format gives reasons to believe that this will change to the better.

6. REFERENCES

- [1] CCSDS (Consultative Committee for Space Data Systems): *XML Formatted data Unit (XFDU) structure and Construction Rules*, CCSDS 661.0-B-1, Blue book, <http://public.ccsds.org/publications/archive/661x0b1.pdf>, 2008.
- [2] Christensen, S. S. *Archival Data Format Requirements*, The Royal Library, Copenhagen, Denmark, The State and University Library, Aarhus, Denmark. http://netarchive.dk/publikationer/Archival_format_requirements-2004.pdf, 2004.
- [3] Cohen, M., Garfinkel, S., Schatz, B. Extending the advanced forensic format to accommodate multiple data sources, logical evidence, arbitrary information and forensic workflow, In: Digital Investigation - The International Journal of Digital Forensics & Incident Response, no. 6, pp. 57-68, 2009.
- [4] Cramer, T., Kott, K. Designing and Implementing Second Generation Digital Preservation Services: A Scalable Model for the Stanford Digital Repository, In: *D-Lib Magazine*, vol. 16, no. 9/10. 2010.
- [5] Enders, M.: A METS Based Information Package for Long Term Accessibility of Web Archives, In: *Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria*, 2010.
- [6] Gillesse, R., Rog, J., Verheusen, A. Life Beyond Uncompressed TIFF: Alternative File Formats for Storage of Master Image File.' In: *Proceedings of the IS&T Archiving Conference*, Bern, Switzerland, 2008.
- [7] ISO 28500:2009, *Information and documentation -- WARC file format*, retrievable via http://www.iso.org/iso/iso_catalogue.htm, 2009.
- [8] ISO 9660:1988, ECMA-119, *Optical media disk imaging format*, retrievable via http://www.iso.org/iso/iso_catalogue.htm, 2010.
- [9] Kejser, U.B.: Preservation copying of endangered historic negative collections, In: Proceedings of the IS&T Archiving Conference, Bern, Switzerland, pp. 177-182, 2008.
- [10] Kim, Y. Digital forensics formats: seeking a digital preservation storage format for web archiving. In *Proceedings of 7th International Digital Curation Conference*, Bristol, United Kingdom, 2011.
- [11] Lavoie, B., Dempsey, L.: Thirteen Ways of Looking at ... Digital Preservation, In: *D-Lib Magazine* vol. 10 no. 7/8, 2004.
- [12] Local Digital Format Registry (LDFR), *File Format Guidelines for Preservation and Long-term Access*, Version 1.0, Library and Archives Canada, <http://www.collections.canada.gc.ca/digital-initiatives/012018-2210-e.html>, 2010.
- [13] Library of Congress. Sustainability of Digital Formats: Planning for Library of Congress Collections, <http://www.digitalpreservation.gov/formats/sustain/sustain.shtml>
- [14] National Archives (UK). *Selecting File Formats for Long-Term Preservation*, http://www.nationalarchives.gov.uk/documents/selecting_file_formats.rtf, 2003.
- [15] Oury, C., Peyrad, S.: From the World Wide Web to digital library stacks: preserving the French web archives In: *Proceedings iPRES 2011, 8th International Conference on Preservation of Digital Objects, Singapore, Singapore*, 2011
- [16] Phillip Katz, *Computer Software Pioneer*, 37, In: The New York Times, May 1st 2000, <http://www.nytimes.com/2000/05/01/us/phillip-katz-computer-software-pioneer-37.html>, 2000.
- [17] PREMIS Editorial Committee, *PREMIS Data Dictionary for Preservation Metadata*, version 2.1, <http://www.loc.gov/standards/premis/v2/premis-2-1.pdf>, 2011.
- [18] Rauch, C., Krottmaier, H. Tochtermann, K.: File-Formats for Preservation: Evaluating the Long-Term Stability of File-Formats, In: *Proceedings ELPUB2007 Conference on Electronic Publishing, Vienna, Austria*, 2007.
- [19] Reich, V., Rosenthal, D. S. H.: Distributed Digital Preservation: Private LOCKSS Networks as Business, Social, and Technical Frameworks, In: *Library Trends*, vol. 57, no. 3, pp. 461-475, 2009
- [20] Rosenthal, D. S. H., Robertson, T., Lipkis, T., Reich, V., Morabito, S.: Requirements for Digital Preservation Systems, A Bottom-Up Approach, In: *D-Lib Magazine*, vol. 11, no. 11, 2005.
- [21] Seadle, M.: Archiving in the networked world: LOCKSS and national hosting, In: *Library Hi Tech*, vol. 28, Issue 4, pp. 710-717 (2010)
- [22] The InterPARES 2. Project. *General Study 11 Final Report: Selecting Digital File Formats for Long-Term Preservation*, [http://www.interpares.org/display_file.cfm?doc=ip2_file_formats\(complete\).pdf](http://www.interpares.org/display_file.cfm?doc=ip2_file_formats(complete).pdf), 2006.
- [23] WARC, Web ARCHive file format, in: Sustainability of Digital Formats Planning for Library of Congress Collections, <http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml>
- [24] Wright, R., Miller, A., Addis, M.: The Significance of Storage in the “Cost of Risk” of Digital Preservation, In: *The International Journal of Digital Curation*, vol. 4, issue 3, pp. 105-122, 2009.
- [25] Zierau, E. *A Holistic Approach to Bit Preservation*, Doctoral Thesis, University of Copenhagen, http://www.diku.dk/research/phd-studiet/phd/thesis_20111215.pdf/, 2011.

Rethinking authenticity in digital art preservation

Perla Innocenti

CCA, University of Glasgow

8 University Gardens

Glasgow, UK, G12 8QH

perla.innocenti@glasgow.ac.uk

ABSTRACT

In this paper I am discussing the repositioning of traditional conservation concepts of historicity, authenticity and versioning in relation to born digital artworks, upon findings from my research on preservation of computer-based artifacts. Challenges for digital art preservation and previous work in this area are described, followed by an analysis of digital art as a process of components interaction, as performance and in terms of instantiations. The concept of dynamic authenticity is proposed, and it is argued that our approach to digital artworks preservation should be variable and digital object responsive, with a level of variability tolerance to match digital art intrinsic variability and dynamic authenticity.

Categories and Subject Descriptors

H.1.1 [Systems and Information Theory]: Value of information.
J.5 [Arts and Humanities]: Arts, fine and performing

General Terms

Documentation, Theory, Verification.

Keywords

Digital preservation. Digital art. Authenticity. Instantiations. Performances. Music notation.

1. DIGITAL CASUALTIES: CHALLENGES FOR DIGITAL ART PRESERVATION

Born digital art is fundamentally art produced and mediated by a computer. It is an art form within the more general “media art” category [1] and includes software art, computer-mediated installations, Internet art and other heterogeneous art types.

The boundaries of digital art are particularly fluid, as it merges art, science and technology to a great extent. The technological landscape in which digital art is created and used challenges its long term accessibility, the potentiality of its integrity, and the likelihood that it will retain authenticity over time. Digital objects – including digital artworks – are fragile and susceptible to technological change. We must act to keep digital art alive, but there are practical problems associated with its preservation, documentation, access, function, context and meaning. Preservation risks for digital art are real: they are technological but also social, organisational and cultural [2].

Digital and media artworks have challenged “traditional museological approaches to documentation and preservation because of their ephemeral, documentary, technical, and multi-part nature” [3]. The technological environment in which digital art lives is constantly changing, and this fast change makes it very difficult to preserve this kind of artwork. All art changes. And these changes can occur at art object level and at context level. In

most circumstances this change is very slow, but in digital art this isn’t the case anymore because it is happening so quickly, due to the pace of technological development.

Surely the increased pace of technological development has more implications than just things happening faster. Digital art, in particular, questions many of the most fundamental assumptions of the art world: What is it a work of art in the digital age? What should be retained for the future? Which aspects of a given work can be changed and which must remain fixed for the work to retain the artist’s intent? How do museums collect and preserve? Is a digital work as fragile as its weakest components? What is ownership? What is the context of digital art? What is a viewer? It is not feasible for the arts community to preserve over the centuries working original equipment and software. And industry has no incentive to reproduce old parts or to make current parts backwards compatible. Furthermore, as Richard Rinehart noted, due to lack of formal documentation methods and the goal to bypass traditional art world’s values and practices, media art works are “becoming victims to their own volatile intent” [4]. Museums have long played a critical role in the creation and transmission of knowledge, culture and identity [5]. As they undergo a metamorphosis from the physical to the virtual, museums continue to serve this custodial role, although their nature and reach might be very different in the future. In particular, as museums invest in collecting digital works, they come to recognize that these works are fragile and may require substantial continued investment in finance and effort to keep them accessible over time.

2. LONG-TERM ACCESSIBILITY OF DIGITAL ART: PREVIOUS WORK

Digital art may seem less physical than traditional art. But as novelist Bruce Sterling noted, “very little materiality, is very, very far from no materiality at all” [6]. The bitstream might be composed by numbers, but the device – the computer – has similar conservation problems as a painting (e.g. humidity, heat, physical damage), plus a whole set of new ones.

Digital preservation is not only about keeping the bits that we use to represent information, but to keep these bits *alive*, as an ongoing activity to ensure recurring value and performance of digital objects, including digital artworks. As Seamus Ross clarified, digital preservation is about “maintaining the semantic meaning of the digital object and its content, about maintaining its provenance and authenticity, about retaining its interrelatedness, and about securing information about the context of its creation and use” [7]. Conservation and restoration are relevant, however they are part of a larger group of activities to ensure longevity for digital objects: collection and repository management, selection and appraisal, destruction, risk management, preserving the context, interpretation and functionality of objects, ensuring a collection’s cohesion and interoperability, enhancing, updating

and annotating, scalability and automation; storage technologies and methods.

In the last decades, much work has been done towards establishing the long-term accessibility of electronic, media and digital art, as well as documenting media and digital art in order to keep it accessible in the future. Some of the key projects and initiatives in this area were started already in the 1970s (for example, the Electronic Art Intermix [EAI] and the Netherlands Media Art Institute [NIMk], Montevideo/Time Based Arts) and further initiatives developed through the following decades, including V2, Matters in Media Art, Forging the Future and DOCAM [8].

These projects and initiatives have contributed to raising awareness on some of the challenges of digital art preservation, examine media and digital art works, explore some specific documentation aspects, and initiate collaborations with other institutions. Nevertheless, much of this work has been survey-like and not particularly well-founded from either a theoretical or methodological perspective. So far, the theoretical aspects of the problem of digital art preservation and curation have been examined without much grounding particularly in experimentation, and not responding to the theoretical and methodological dilemmas posed by digital art (e.g. transience, emergence, and lack of fixity). Also the long term preservation of documentation for digital art has not yet been systematically addressed. Documentation for digital art is at risk as much as digital artworks themselves, and needs sustainable business and organisational models to be preserved in the long term.

It is evident that digital art is a new phenomenon that requires a new suite of methodologies.

3. MY INVESTIGATION

The goal of the research project *Preserving Computer-Generated Imagery: Art Theory, Methods and Experimental Applications* [9] that I am conducting at the University of Glasgow is to contribute to laying the foundations for a preservation framework of digital art and identifying interdisciplinary synergies with areas such as digital preservation, philosophy of art, archival science and information management. Digital art is after all data designed to be constructed (represented, viewed, experienced) in particular ways, whose theoretical implications need consideration. The methodology that I have chosen to take is bottom up, to try to understand how digital art works. That is: I am starting with the works, the conservators and the creators, using a mixed method of humanistic, social science [10] and engineering approaches. So I have decided to adopt a two-step method: onsite visits to major international collectors of digital art and in-depth interviews with their staff; and experimentation with testbeds to assess preservation methods and processes. I am using a mixed method of humanistic, social science and engineering approaches.

The humanistic element of it is the art history aspect, and the reflection on what is a work of art in the digital age and what is the context of digital art. I am presenting some reflections on authenticity and longevity for digital art in section 4, ideas which have been further shaped by my social science approach. From a social science perspective I have visited and talked with some of the most important collectors of digital art conducting a whole series of interviews, which have provided me a window on the practices of different organisations working with digital art. I have borrowed methods from anthropology and grounded theory. In my first phase of ethnographic process of observation of digital

media art, I looked at key digital art organizations and how they are collecting, curating, preserving, displaying, and financing digital art. I conducted onsite in-depth interviews, visits and observations because what I am told is sometimes at variance with what is being done. The organizations that I targeted and selecting for my case studies are major international collectors of digital artworks and digital art documentation. I visited ZKM | Media Museum at the ZKM | Centre for Art and Media (Germany), Ars Electronica Centre – AEC (Austria), The Hirshhorn Museum and Sculpture Garden, (USA), Smithsonian American Art Museum and Lunder Conservation Center (USA), Museum of Modern Art in San Francisco – SFMOMA (USA), Berkeley Art Museum – BAM (USA), Museum of Modern Art – MOMA (USA), Whitney Museum (USA), and NIMk (The Netherlands). The complexity of maintaining the object longevity and the myriad of change that can occur over time means that we need to talk with organizations that have decades of experiences to understand what needs to be done in this area. Interviews with stakeholders of digital art preservation (museum directors, conservators, curators, registrars, technicians) are a new approach in this area. I also conducted interviews and observations with selected digital artists (John Gerrard, Studio Azzurro, Maurice Benayoun) for an additional analysis of relevant aspects of preservation for digital artworks.

4. REFLECTIONS ON AUTHENTICITY FOR DIGITAL ART

Two aspects emerged from the first phase of my investigation strike me as key for digital art preservation: the intrinsic performing nature of digital art, and the dynamic nature of digital art authenticity.

4.1 Digital art as a process of components interaction

The ability to establish authenticity in a digital object is crucial for its preservation [11]. Even if the concept of authenticity is highly nuanced in the digital age, it is still a starting point for discussion about digital art. But to talk about authenticity we need to look at how digital art is created and rendered. For example, the image of the work *Bubbles* (2001) by Muench and Furukawa in the ZKM | Media Museum, is a process of interaction of many components: for this example particularly, the file in which the data matrix representing the image is stored, and the software capable of interpreting and rendering this data form. If we were to explore this example in full, we would also need to discuss the hardware, the data projector, the screen, and the relationships (including intended effects) that all this has with the viewer.

4.2 Digital art as performance

This interaction of components leads me to think that all digital art is a performance, and more than a performance between the viewer and the object. In this particular instance, the performance that I am actually talking about is the *performance of the work*. Because a digital artwork consists of a set of code, and for the artwork *to become*, it must be performed. Before the viewer interacts with the digital artwork, this process of becoming has to occur. For example in the case of John Gerrard's 3D real time work *Grow Finish Unit (near Elkhart, Kansas)* (2008) at the Hirshhorn Museum, the algorithm developed by Gerrard needs to be performed in order for the work itself – the real time 3D – to come to life.

This problem isn't actually unique to digital art. For example, within the AktiveArchive project, Johanna Phillips and Johannes Gfeller wrote interesting reflections about reconstruction and well-informed re-performances of video art [12]. But in the field of digital art, it is nearly another construct. Some very groundbreaking work in the documentation of performances has been done by Richard Rinehart, former digital media artist and director of the UC Berkeley Art Museum/Pacific Film Archive. Rinehart produced a promising theoretical approach based on a formal notation system for digital and media art creation, documentation and preservation: the Media Art Notation System (MANS) [13]. He compared media art to the performative arts, because media art works do not exist in a stable medium, and are inherently variable and computational. Their preservation is thus an interpretive act. Given the similar variability of music and media arts, Rinehart considers as appropriate a mechanism like a musical score for binding the integrity of media art works apart from specific instruments.

4.3 Instantiations, authenticities and documentation of digital art

Considering digital art as performance leads to some interesting reflections about its instantiations. As Seamus Ross observed, the "first renderings of digital objects might best be referred to as an initial 'representation or instantiation' (II). The problem is: how can we record the functionality and behaviour as well as the content of that initial instantiation (II) so that we can validate subsequent instantiations? Where subsequent instantiations (SI) share precision of resemblance in content, functionality, and behaviour with the initial instantiations, the 'SIs' can be said to have the same authenticity and integrity as the 'IIs'" [14]. This notion of precision of resemblance is intended to reflect the fact that initial instantiations of digital objects and subsequent ones will not be precisely the same, but will have a degree of sameness. This degree of sameness will vary overtime – in fact in the case of digital objects it is likely to decline as the distance between the initial instantiation and each subsequent one becomes greater, although this degree of variation may be mitigated by such circumstances as for example the frequency at which the digital object is instantiated. So each time a digital work of art is instantiated, it has a greater or lesser precision of resemblance to the initial instantiation, which the artist created. The subsequent instantiations represent with greater or lesser degrees of accuracy the intentionality of the artist. Whether they have greater or lesser degrees of authenticity is a separate but fundamentally important question and need to be considered in the context of, for example, the authenticity of performances. The UNESCO Guidelines for the Preservation of Digital Heritage mentions the question of assessing an acceptable level of variance of such instantiations [15]. This was also more recently highlighted by Richard Rinehart, in relation to the ecological balance of changes in the technological environment of digital art [16].

The intrinsic performing nature of digital artworks makes them allographic rather than autographic works, along the distinction described by Nelson Goodman [17]. So I would like to draw a parallel between the instantiation of the code in a digital work, and the instantiation of the notation in a music performance, as described by John Butt and Dennis Dutton.

We often assume that music notation is a rigid set of instructions. In reality, sometimes notation is the result of performance,

sometimes it is a reminder, and sometimes it is just an example. There is no single process from notation to performance. The notation is going in all directions, with a complex relationship between sender and receiver. In his seminal book *Playing with history: the historical approach to musical performance* [18], John Butt has questioned whether "authenticity" is still an appropriate term for music performance given that, in performance terms, it tends to condemn its negative to a sort of fake status. In music, partly through Butt's effort, we now tend to use the term "historically informed performance". In his reflection on nominal authenticity in the arts, Dutton writes, "the best attitude towards authenticity in music performance is that in which careful attention is paid to the historic conventions and limitations of a composer's age, but where one also tries to determine the artistic potential of a musical work, including implicit meanings that go beyond the understanding that the composer's age might have derived from it" [19].

The dynamic notion of authenticity of digital art might seem to be in contrast with the notion of material authenticity that has been constructed for historical artworks. If we look at authenticity in object conservation in museums, authenticity is a term associated with the original material components and process in an object, and its authorship or intention. For example, in his critique of traditional conservation ethics, Jonathan Kemp describes "authenticity in the sense of 'original material', traditionally one aspect of an object charged with the assignation of a 'truth value' that legitimizes some aesthetic experiences" [20]. However these conservation principles are socially constructed processes mediated by technology-based practices, whereas the object keeps changing: it deteriorates, its context might change, and the way that it is conserved and re-displayed will change. The role of conservators and of museums also changes over time. Therefore the conservators are caught between reconciling fidelity to the original artist intention, and fidelity to the passage of time. Joseph Grigely also argued that any work of art is subject to a "continuous and discontinuous transience" [21], that is integral to its authenticity. This means that any work of art – I shall add including digital art – is not fixed in a single point in time, but it is rather in a "continuous state of becoming", as Heather MacNeil and Bonnie Mak elegantly pointed out [22]. Like in Penelope's tale, conservators are actively constructing and reconstructing the authenticity of a work based on their understanding of its nature and the current conventions and assumptions for conserving it. These reflections on instantiations and authenticity led my attention to the concept of authenticity in electronic records. As Jennifer Trant noted, "archives have been challenged to manage electronic records as evidence for several decades [...]" [23]. Like art conservators, archivists and record keepers are concerned with issues of fidelity. The trustworthiness of a record rests primarily on its fidelity to the original event, from which the record arises. The concept of provenance – a well-documented chain of custody – is thus a fundamental archival principle, which helps establishing authenticity [24]. This has parallels with my reflections on instantiations of digital artworks. If we look at computer-based art from the point of view of performance and archival authenticity, what is then really important is a trustworthy chain of documentary evidence about the work genuine origins, custody, and ownership in the museum collection. Authenticity is not an original condition, but it is rather a dynamic process. Digital artworks are pushing the boundaries of traditional conservation practices and the notion of historicity. For

example, let's look at the ongoing preservation strategy devised within the Digital Art Conservation project [25] for the interactive media art work *The Legible City*, 1989-1991 in the ZKM | Media Museum. This strategy could be seen as the equivalent of rewriting an older music score to adapt it to a modern or different instrument. On one hand, this iconic interactive installation is based on proprietary, work-specific software; on the other, it uses obsolete hardware and custom-made components. Such combination makes the preservation of *Legible City* a costly and risky business, both for the price of maintaining its Indigo 2 computer (no longer produced by Silicon Graphics) and because of the potential weak point represented by its specially-built analog-digital transformer. Conservators at ZKM examined, documented and created a fully-functional replica of this transformer (the interactivity intended as part of the installation was also recorded), and software porting to another operating system is currently being evaluated by the ZKM as a more sustainable long-term preservation solution for the Indigo 2 computer. Some conservators and curators might argue that the replacement of the historical software and transformer challenges the historicity and originality of the artwork. However, digital art collectors need to come to terms with the fact that it will not be possible to guarantee forever original working equipment: in order to be kept alive, digital artworks will need to be adapted to a new technology [26]. This artwork at ZKM is in the state of *becoming*. This idea of becoming is clearly referenced in the work of Heather McNeil Bonnie and Mak about constructions of authenticity, and this goes back to the notion that digital art becomes, which I mentioned earlier. Digital works are in a state of evolution.

5. CONCLUSIONS

With this paper, I hope to stimulate discussions about current and future approaches for digital art preservation, and contribute to the interdisciplinary foundations of a scientific framework for digital art preservation.

Authenticity – as MacNeil and Mak clearly pointed out – is a social construct, whose parameters and contents are always changing and under negotiation. Authenticity allows us to author stability in our disciplines. The current fast-paced digital environment defies the traditional structures of stability that have been authored for traditional art. Therefore our approach to digital artworks should be variable and digital object responsive, with a level of variability tolerance to match digital art intrinsic variability and dynamic authenticity, as outlined in this paper. The designated community for whom we are preserving should also be identified, together with the modality of restaging digital works and of preserving the related digital documentation. In conclusion, if conservation for digital art is a moving target, then our scientific methodology should be a moving gun.

6. ACKNOWLEDGMENTS

I am deeply indebted to Prof. Seamus Ross at the Faculty of Information, University of Toronto, for his precious suggestions, guidance and support throughout this research, and more recently to Prof. John Butt at the University of Glasgow, for sharing his knowledge and experience on musical performance. I am also very grateful to all my interviewees for the time and helpful insights that they have shared with me regarding conservation and preservation for digital art.

7. REFERENCES

- [1] Christiane Paul, *Digital art*, 2. ed., Thames & Hudson, 2008; id. (ed.), *New Media in the White Cube and beyond. Curatorial Models for Digital Art*, C. University of California Press, Berkeley, 2008; Alain Depocas, Jon Ippolito, and Caitlin Jones (eds.), *Permanence Through Change: The Variable Media Approach*, Guggenheim Museum, New York and The Daniel Langlois Foundation for Art, Science & Technology, Montreal, 2003; Oliver Grau and Rudolf Arnheim (eds.), *MediaArtHistories*, MIT Press, Cambridge, 2007; Wolf Lieser, *Digital Art. Neue Wege in der Kunst*, h.f. ullmann, Berlin, 2010.
- [2] Perla Innocenti, Andrew McHugh, and Seamus Ross, "Tackling the risk challenge: DRAMBORA Interactive", in: Paul Cunningham and Miriam Cunningham (eds.), *Collaboration and the Knowledge Economy: Issues, Applications, Case Studies*, IOS Press, Amsterdam, 2008.
- [3] Richard Rinehart, "The Media Art Notation System: Documenting and Preserving Digital/Media Art", in: *Leonardo: Journal of the International Society for the Arts, Sciences and Technology*, vol. 40, no. 2, 2007, p. 181; available online at: <http://www.coyoteyp.com/rinehart/papers.html>.
- [4] Rinehart 2007, p. 181.
- [5] Bennett, T. *The Birth of the Museum. History, Theory, Politics*, Routledge, London, New York, 2009; Knell, J.K., MacLeod, S., Watson, S. (eds.), *Museum Revolutions. How Museums Change and Are Changed*, Routledge, London, New York, 2007; Graham, B. Cook, S. *Rethinking Curating. Art after New Media*. Leonardo, MIT Press, 2010 and Altshuler, B. *Collecting the new: a historical introduction*, in *Collecting the New: Museums and Contemporary Art*, Princeton University Press, 2007.
- [6] Sterling, B. "Digital Decay", in: Alain Depocas, Jon Ippolito, and Caitlin Jones (eds.), *Permanence Through Change: The Variable media Approach*, Solomon R. Guggenheim Museum, New York, Daniel Langlois Foundation, Montreal, 2003, pp. 10–22, http://www.variablemedia.net/e/preserving/html/var_pub_index.html.
- [7] Ross, S. *Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries*, Keynote Speech at the European Conference on Research and Advanced Technology for Digital Libraries (ECDL) 2007, Budapest, Hungary, 17 September 2007, p. 2, www.ecdl2007.org/Keynote_ECDL2007_SROSS.pdf.
- [8] Electronic Art Intermix (EAI), www.eai.org/index.htm; Netherlands Media Art Institute NIMk, Montevideo/Time Based Arts, www.nimk.nl/. Further projects and initiatives developed over the last decades are: Independent Media Arts Preservation (IMAP), since 1999, www.imappreserve.org/; International Network for Conservation of Contemporary Art (INCCA), since 1999, www.incca.org/; Variable Media Network, 2000-2004, www.variablemedia.net/; AktiveArchive Project, 2001-2009, www.aktivearchive.ch/content/projekte.php; Archiving the Avant-Garde: Documenting and Preserving Variable Media Art, 2002-2010,

- www.bampfa.berkeley.edu/about/avantgarde; 404 Object Not Found. What remains of Media Art?, 2003. Sadly this project is no longer available online. A project description is at <http://nimk.nl/eng/404-object-not-found-what-remains-of-media-art>; V2_ Capturing Unstable Media, 2003, <http://capturing.projects.v2.nl/>; Matters in Media Art: collaborating towards the care of time-based media, since 2003, www.tate.org.uk/about/projects/matters-media-art; packed.be, since 2003www.packed.be/; PANIC (Preservation webservices Architecture for New media and Interactive Collections), since 2003; this project website is being preserved by the National Library of Australia at <http://pandora.nla.gov.au/tep/49720>; Inside Installation Project, 2004-2007, www.inside-installations.org/home/index.php; 40yearsvideoart.de, 2004-2006, www.40jahrevideokunst.de/main.php?p=3; Ludwig Boltzmann Institut - Medien.Kunst.Forschung, 2005-2009, <http://media.lbg.ac.at/de/index.php>; Forging the Future: New Tools for Variable Media Preservation, 2007-2008, <http://forging-the-future.net/>; DOCAM - Documentation and Conservation of the Media Arts Heritage project, 2005-2009, <http://www.docam.ca/>.
- [9] Some aspects of my research have been published in Perla Innocenti, "Theories, methods and testbeds for curation and preservation of digital art", in: IS&T Archiving 2010 Preservation Strategies and Imaging Technologies for Cultural Heritage Institutions and Memory Organisations Conference, 1-4 June 2010, The Hague, pp. 13-17.
- [10] Woolgar, S., 'Technologies as cultural artefacts, in: Dutton, W. (Ed.), Information and Communication Technologies - Visions and Realities, Oxford University Press, 1996.
- [11] Ross, S. "Position Paper on Integrity and Authenticity of Digital Cultural Heritage Objects," in: Digicult. Integrity and Authenticity of Digital Cultural Heritage Objects, Thematic Issue 1, August 2002, pp. 6-8, www.digicult.info/downloads/thematic_issue_1_final.pdf.
- [12] Phillips, J. "The reconstruction of video art. A fine line between authorised re-performance and historically informed interpretation," in: Irene Schubinger (ed.), Reconstructing Swiss Video Art from the 1970s and 1980s, JRP Ringier, Zurich, 2009, pp. 158-165; and Johannes Gfeller, "The reference hardware pool of AktiveArchive at Bern University of Arts. A basis for a historically well-informed re-performance of media art," in: Schubinger 2009, pp. 166-174. Some useful reflections are also published in: Erma Hermens, and Tina Fiske, T. (eds.), Art Conservation and Authenticities. Material, Concept, Context, Proceedings of the international conference held at and in Howard Besser, "Longevity of Electronic Art," in: David Baerman, and Franca Garzotto (eds.), International Cultural Heritage Informatics Meeting: Cultural Heritage and Technologies in the Third Millennium, Proceedings from the ichim01 Meeting, Milan, Italy, September 3-7, 2001, vol. 1, Archives & Museum Informatics, Pittsburgh, 2001, pp. 263-275, www.archimuse.com/publishing/ichim_01_TOC1.html.
- [13] Rinehart 2007.
- [14] Ross, S. "Approaching Digital Preservation Holistically," in: Tough, A. and Moss, M. (eds.), Record keeping in a hybrid environment : managing the creation, use, preservation and disposal of unpublished information objects in context, Chandos Press, Oxford, 2006, pp. 115-153.
- [15] UNESCO (National Library of Australia), Guidelines for the Preservation of Digital Heritage, Report, 2003, § 16.7, <http://unesdoc.unesco.org/images/0013/001300/130071e.pdf>.
- [16] Innocenti, P. Interview on curation and digital preservation of time-based/media art of with Richard Rinehart, Berkeley Art Museum (BAM), 25 March 2010).
- [17] Goodman, N. Languages of Art: an Approach to a Theory of Symbols, Oxford University Press, London, 1969 (in particular the chapter on Art and Authenticity and on The Unfakable).
- [18] Butt, J. Playing with History: the Historical Approach to Musical Performance, Cambridge University Press, Cambridge, 2002.
- [19] Dutton, D. "Authenticity in Art," in: Jerrold Levinson (ed.), The Oxford Handbook of Aesthetics, Oxford University Press, New York, 2003, www.denisdutton.com/authenticity.htm.
- [20] Kemp, J. "Practical Ethics v2.0," in: Richmond, A. (ed.), Conservation. Principles, Dilemmas and Uncomfortable Truths, Butterworth-Heinemann, London, Amsterdam, Boston, Heidelberg et al., 2009, pp. 60-72.
- [21] Grigely, J. Introduction, in Texualterity: Art, Theory and Textual Criticism, University of Michigan Press, Ann Arbor, 1995, p. 1.
- [22] MacNeil, H. and Mak, B. "Constructions of Authenticity," in: Library Trends, vol. 56, no. 1: Preserving Cultural Heritage, Summer 2007, pp. 26-52.
- [23] Trant, J. "Emerging Convergence? Thoughts on museums, archives, libraries and professional training," in: Museum Management and Curatorship, vol. 24, no. 4, Dec. 2009, pp. 369-387.
- [24] In archives authenticity is "the quality of being genuine, not counterfeit, and free from tampering, and is typically inferred from internal and external evidence, including its physical characteristics, structure, content, and context." see: The Society of American Archivists (SAA), A Glossary of Archival and Records Terminology, available online at: www.archivists.org/glossary/term_details.asp?DefinitionKey=9. In terms of evidence, "provenance is a fundamental principle of archives", defined as "information regarding the origins, custody, and ownership of an item or collection." See: The Society of American Archivists (SAA), A Glossary of Archival and Records Terminology, available online at: www.archivists.org/glossary/term_details.asp?DefinitionKey=196.
- [25] Digital Art Conservation, 2011, ZKM | Center for Art and Media Case Study: Jeffrey Shaw, *The Legible City*. <http://www02.zkm.de/digitalartconservation/index.php/en/exhibitions/zkm-exhibition/nnnnnjeffrey-shaw.html>.
- [26] Innocenti, P. Interview on digital preservation on media art of with Dr. Bernhard Serexhe, ZKM | Media Museum, Karlsruhe, 12 August 2008.

Describing Digital Object Environments in PREMIS

Angela Dappert
Digital Preservation Coalition
% The British Library
96 Euston Road
London, NW1 2DB, UK
angela@dpconline.org

Sébastien Peyrard
National Library of France
Bibliographic and Digital Information Department
Quai François-Mauriac
75706 Paris Cedex 13
sebastien.peyrard@bnf.fr

Janet Delve
The School of Creative Technologies
The University of Portsmouth
Winston Churchill Avenue
Portsmouth, PO1 2DJ
janet.delve@port.ac.uk

Carol C.H Chou
Florida Digital Archive
Florida Virtual Campus
5830 NW 39th Ave.
Gainesville, FL 32606
U.S.A.
cchou@ufl.edu

ABSTRACT

“Digital preservation metadata” is the information that is needed in order to preserve digital objects successfully in the long-term so that they can be deployed in some form in the future. A digital object is not usable without a computing environment in which it can be rendered or executed. Because of this, information that describes the sufficient components of the digital object’s computing environment has to be part of its preservation metadata. Although there are semantic units for recording environment information in PREMIS 2, these have rarely, if ever, been used. Prompted by increasing interest in the description of computing environments, this paper describes on-going efforts within the PREMIS data dictionary’s Editorial Committee to define an improved metadata description for them.

Keywords

H.1.0 [General Models and Principles]: PREMIS; preservation metadata; technical environments; software preservation; hardware preservation; representation information; representation information network; conceptual modelling.

1. INTRODUCTION

“Metadata” is information about an object that is needed in order to manage that object. “Digital preservation metadata” is the information that is needed in order to preserve digital objects successfully in the long-term so that they can be deployed in some form in the future [1]. A digital object is not usable without a computing environment in which it can be rendered or executed. Digital objects are normally not self-descriptive and require very specific intermediary tools for access by humans and specific knowledge for interpreting them. Neither may be commonly available amongst a repository’s Designated Community (as defined in OAIS [2]). Because of this, information that describes the sufficient components of the digital object’s environment constitutes essential representation information that is needed in order to be able to use the digital object and to make it understandable in the future.

Core metadata for the digital preservation of any kind of digital object is specified in the PREMIS Data Dictionary [3], a de-facto standard. Core metadata is the metadata that is needed by most preservation repositories, rather than application or

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

content specific metadata defined for niche uses. Metadata about digital objects’ computing environments must be preserved together with the digital objects as part of their core metadata.

In addition to describing an Object’s representation information, some computing environments, such as software, can themselves be the primary objects of preservation, as may be the case for computer games. They may also take the role of a software Agent in a preservation Event, and may require a thorough metadata description for those reasons.

Although there are semantic units for recording environment information in PREMIS version 2, these have rarely, if ever, been used. In 2011, the PREMIS data dictionary’s Editorial Committee commissioned a working group to re-examine what computing environment metadata needs to be captured in order to be able to successfully redeploy digital objects in the long-term. This paper describes these on-going efforts. The result may be implemented in version 3 of the PREMIS Data Dictionary.

2. PRESERVING COMPUTING ENVIRONMENTS

2.1 The Current State

In version 2 of the PREMIS Data Dictionary [3], there are four key entities that need to be described to ensure successful long-term preservation of digital objects: Object, Event, Agent and RightsStatement. The Object entity provides two places to describe subordinate environments. For one, there is the “environment” semantic unit that permits the description of software, hardware and other dependencies. Rather than being an entity per se, an Environment is modelled as a semantic unit container that belongs to an Object and is, therefore, subordinate to the Object entity. The second environment-related semantic unit is the “creatingApplication” that also is sub-ordinate to the Object entity. Creating applications are outside the scope of an OAIS repository and have therefore been historically treated separately from other Environment descriptions. In a generic digital preservation framework that is not restricted to OAIS use, but supports the end-to-end digital preservation life-cycle, one would describe Environments uniformly, no matter in what context they are used. Our proposal prefers a solution that accommodates this view.

Its subordinate position to Objects means that Environments can only be captured to describe an Object’s computational context. This has the following limitations:

- Environments are too complex to be handled in an Object repository.

- Environments are rarely specific to a single Object, resulting in their redundant spread across different Objects. This results in
 - unnecessary verbosity;
 - cumbersome management of Environment descriptions as they evolve.
- They are unable to describe stand-alone Environments and unable to be used for modelling an Environment registry that describes Environment components without the need for creating Objects.
- They are primarily applicable to computing environments and do not include representation information in the broader sense. This restricts the description to a technical level rather than to a level that comprehensively enables redeployment.

Our use case analysis identified the five desirable relationships illustrated in Figure 1. Because Environments are subordinate to Objects, it is impossible to express the latter four of them.

1. An Object specifies its Environment, i.e. its computational context. This is the existing relationship in PREMIS 2.
2. An environment (for example, software source code) is to be preserved as first-class entity in its own right. It is described as Environment and takes on the role of an Object.
3. An environment is described as Environment and takes the role of an Agent (for example, as software Agent involved in a preservation action Event).
4. An environment is described as Environment and is related to another Environment through inclusion, dependency, derivation or other relationships.
5. An environment is described as Environment and has an Event associated with it (for example, a creation or versioning Event).

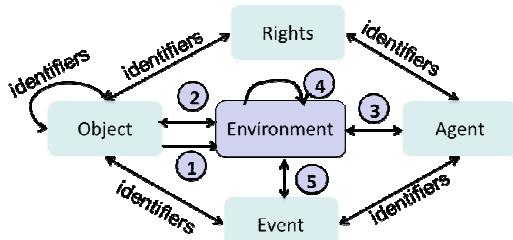


Figure 1: The basic entities of the PREMIS Data Dictionary (in blue) with the desired Environment entity and their relationships.

Another limitation is that in PREMIS 2, Environments are unable to refer to external dedicated registries, which would enable the delegation of "up-to-date and complete" information to an external source if needed. The identified shortcomings may be the reason that the Environment semantic container in PREMIS is rarely used.

The goal of the PREMIS Environment Working group is to rethink the metadata specification for environments. Their description must meet the improved understanding of how to ensure their longevity.

2.2 Related Work

The comprehensive conceptual model of the digital preservation domain in Dappert and Farquhar [4] includes Environ-

ments, Requirements (including significant characteristics) and Risks as first-order entities and justifies why this is beneficial.

There are also several efforts in the digital preservation community to specify the metadata needs for certain aspects of computing environments.

Specialised metadata has been defined to support the preservation of software. For example, "The Significant Properties of Software: A Study" project [5, 6] identified *Functionality, Software Composition, Provenance and Ownership, User Interaction, Software Environment, Software Architecture and Operating Performance* as the basic metadata categories for software that should be applied on *Package, Version, Variant* and *Download* level. The Preserving Virtual Worlds project [7], POCOS [8], SWOP [9] and DOAP [10] have made proposals for software preservation metadata. Examples of software repositories, the NSRL National Software Reference Library [11], MobyGames [12] and AMINET [13] illustrate practically used metadata schemas, but do not necessarily support digital preservation functions. JHOVE [14], PRONOM [15], UDFR [16] and the Library of Congress [17] have defined metadata that is needed to technically or qualitatively describe file formats and have built repositories based on their metadata descriptions. This includes some software metadata specifications, which, for PRONOM, are now available in a linked data representation and for UDFR contains software description in the recently released UDFR database [18].

There are metadata initiatives that address more complex dependencies. The Virtual Resource Description Framework (VRDF) [19] captures virtualized infrastructures; the Cloud Data Management Interface (CDMI) [20] "describes the functional interface that applications use to create, retrieve, update and delete data elements from the Cloud"; and the Web Service Definition Language (WSDL) [21] describes network services as a set of endpoints operating on messages.

The KEEP project on emulation [22] designed a prototype schema for the TOTEM database [23]. It is a recent move towards building a repository for describing the technical properties of computing and gaming environments including software and hardware components. The IIPC [24] has developed a technical database based on a computing environment schema as foundation for web archiving, and TOSEC (short for "The Old School Emulation Centre") [25] "is dedicated to the cataloguing and preservation of software, firmware and resources for microcomputers, minicomputers and video game consoles."

The TIMBUS project [26] addresses the challenge of digital preservation of business processes and services to ensure their long-term continued access. TIMBUS analyses and recommends which aspects of a business process should be preserved and how to preserve them. It delivers methodologies and tools to capture and formalise business processes on both technical and organisational levels. This includes preservation of their underlying software infrastructure, virtualization of their hardware infrastructure and capture of dependencies on local and third-party services and information. This means that, in addition to technical preservation metadata, it draws on metadata standards that capture business processes, such as BPMN [26], and identifies forms of supporting business documentation needed to redeploy processes and services.

Environments correspond to the “Representation Information” of the OAIS information model [2]. Representation Information is “the information that maps a Data Object into more meaningful concepts” [2]. Examples for a specific .docx file would be its file format specification that defines how to interpret the bit sequences, a list of software tools that can render it, hardware requirements, the language in which the contained text is written, and context information that states the author, purpose and time of its writing. Environments include documentation, manuals, underlying policy documents, cheat sheets, user behaviour studies, and other soft aids for interpretation.

3. MODELLING CHOICES

The following principles guided us through the modelling choices:

- Ensure backward compatibility with the existing PREMIS Data Dictionary,
- Ensure compliance with the OAIS information model,
- Provide straightforward Data Dictionary semantics that are easy to implement and that can be implemented within the existing XML Schema and PREMIS ontology,
- Provide clear mapping of historic Environment features to the newly proposed ones.
- Permit an Environment instance to describe a physical item such as software, hardware, a format, a document, a policy or a process. It may or may not be in digital form. It may be more or less concretely specified.

3.1 A Possible Solution

We propose to treat Environments as first class entities that do not have the limitations listed in Section 2.1. Treating Environments as first class entities also makes it more natural to model preservation actions that directly impact Environments, such as data carrier refresh or emulation, as easily as preservation actions that directly impact Objects, say migration. This is particularly important for the preservation of computer games and other kinds of software. While describing those actions is possible with the PREMIS model in version 2, it is not doable in a very natural way.

3.2 Supporting Different Verbosity Needs

Having a dedicated Environment entity gives implementers the ability to make precise and complete descriptions that can be shared with others. To ensure that all needed levels of description can be realised using the PREMIS 3 Data Dictionary, we considered 3 description levels that were designed to match 3 different verbosity levels.

- The most concise: Full outsourcing to an external description. Here the implementer merely wants to point an Object, or an Agent, to a description of its supporting Environment available elsewhere, most likely in some technical registry. This could be achieved by adding a linkingEnvironmentIdentifier from the Objects and the Agents without maintaining the resource that is being referred to.
- The intermediate one: A link is made between an Object or Agent, and its supporting Environment. The Environment instance is described and maintained in the repository, but its components are summarised within its description, rather than elaborated as individual Environments with precise descriptions of all their semantic units that are then linked to each other. This Environment description can be shared

across Agents and Objects, but its component descriptions are not usable individually.

- The most verbose, and precise one: the Environment instance is fully described as a network of modular components, where each Environment is a separate instance. This can be achieved by adding relationships between Environments.

New PREMIS semantic units for Environments should support these description needs, and each more concise verbosity level is built on the basis of the semantic units of the more verbose levels. This way we can maintain a single consistent data dictionary while allowing different levels of description.

3.3 Modelling a Catch-All Term Precisely

Depending on the context, “Environment” can refer to different things. Here are some examples:

- “This operating system only runs on a 64-bit environment”. The environment is hardware, but it is a category consisting of several hardware architectures.
- “This data object can be read on a European NES Games Console environment”. Here the Environment is defined precisely and integrates hardware (including cartridge and controllers) and software (notably the BIOS) at the same time.
- “This ePUBReader plugin requires Firefox 3.0 or later as an execution environment”. Here the Environment merely references software, without pointing to a precise version (all Firefox versions above 3.0 are supposed to work).

These examples demonstrate the following characteristics:

- Environments can connect to other Environments and can consist of related Environment components at lower levels of granularity.
- Depending on the context, as determined by business requirements, different environment subsets are relevant. An Environment can be atomic, freely usable within other Environments; but it can also be a set of running services that achieve a defined purpose (e.g. render an object).
- Environments have a purpose. They allow objects to be rendered, edited, visualised, or executed.
- Some Environments are generic; only the critical aspects of the Environment are specified. Several versions of the Environment or Environments with the same relevant behaviour can be used in its stead.
- Others are specific, real-world instances that are being used or have been used in the lifecycle of preserved Objects.

For capturing the connected nature of Environments, we decided not to introduce a separate concept for “components”. Instead, we treat Environments as entities that can be recursively defined by logical or structural relationships of sets of other Environments. As with other kinds of aggregation, experience proves that, in an implementation-dependent context, what is the top-level entity and what constitutes components varies and results in the choice of different subsets of Environments. Using a recursively-defined Environment entity means that Environments can be flexibly reused in order to create new Environments as dictated by changing business needs. As we had stated that Environments correspond to the “Representation Information” of the OAIS information model, the recursively defined Environment entity forms a Representation Information Network.

3.4 Referring to External Registries

PREMIS evolved from an OAIS tradition. Its goal is to define all preservation metadata that is needed to safeguard Objects stored in an OAIS repository. This excludes events before the Object is ingested into the repository and focuses on the preservation of bitstreams, files and structurally related sets of them-files, captured as representations. It was not intended that it would take the role of a registry, where descriptions and definitions are stored for reuse. Technical registries share with PREMIS the aims of supporting “the renderability, understandability of digital objects in a preservation context” and of representing “the information most preservation repositories need to know to preserve digital materials over the long-term”. Technical registries do NOT describe content “held by a preservation repository”.

As the above examples show, for preservation purposes, an Environment can be a generic description of technical or other characteristics that intend to make the preservation task easier for preservation repositories, but can be increasingly concrete to the point where it would describe a concrete custom-tailored environment for a specific repository. The two domains of registry and repository touch. In a Linked Data implementation there is an almost seamless continuum from the repository preserving digital objects to the external environment descriptions in external registries.

Adding the Environment entity broadens the scope of PREMIS. It focuses no longer only on the Objects preserved in a repository, but also on the representation information needed to render or execute the Object. It captures its reticular nature and core semantics with a new dedicated entity and its semantic units. In the extreme, one could even imagine technical registries using “premis:Environment” natively to describe standalone Environments without relating them to any Object or Agent.

3.5 Matching Environments to the Existing Data Model

We propose to make Environment a new first-class entity so that it can be described with its own semantic units. Therefore, we need to match it to the existing data model, so that backwards compatibility is maintained and so that it is clear when something should be described as an Object, an Agent or an Environment.

In order to achieve reusability and varying levels of specificity an Environment instance should **describe** its characteristics but it should **not state how it is used** in an OAIS repository.

Within an OAIS repository an Environment can take three roles:

- It can take the role of representation information for an Object so that the Object can be redeployed successfully in the future (relationship 1 depicted in Figure 1).
- It can be preserved in the repository for example, to preserve software or a computer game (relationship 2 depicted in Figure 1).
- It can act as an Agent involved in an Event (or, less likely, in a RightsStatement) (relationship 3 depicted in Figure 1).

The fact that an Environment takes on any of these roles is specified in the Object and Agent that captures this information. That is to say that, for example, if an Environment component describes an Agent that is involved in a preservation action Event then a corresponding Agent instance should be created and related to the Environment description. If an Environment

component is to be preserved, then a corresponding Object instance should be created, the Environment’s content has to be captured as an Information Package so that it can be considered an Object, and the instance should be related to the Environment description. If one wishes to merely specify the Environment as representation information for an Object, then again, the Object instance should be created and related to the Environment description.

3.6 Identifying Environments

As indicated in Figure 1, the solution for capturing Environments needs to specify how Environments are to be identified and how other entity instances should link to them. PREMIS 2 offers several different ways of identifying and linking to entity instances. The proposed solution should mirror them for consistency’s sake. The existing approaches include:

- Linking to an entity instance through the **identifier type and value** of the target instance:
linking[Entity]Identifier, to unambiguously link an instance of one entity to an instance of another kind of entity, e.g. an Object to an Event; these links can be particularised with a linking[Entity]Role that allows one to specify the role of the referred entity.
relationship, to unambiguously relate different instances of the same entity, i.e. an Object to another Object. This relationship must be particularised with a type and a subtype. Currently the type values “structural” and “derivation” are suggested values in the Data Dictionary.
dependencyIdentifier, to relate an Object to a file that is needed to support its delivery, e.g. a DTD or an XML Schema.
- Linking to an entity instance through a **registry key: formatRegistryKey**, to relate a file or bitstream Object to a description of its format in an external registry.
- Linking to an entity instance through a **designation: formatDesignation**, to identify a format by name and version.

An Environment as a PREMIS entity must define its identifierType and identifierValue as all other PREMIS entities do. PREMIS Environments are instances that can be linked to from other entities using the premis:identifier mechanism through a linkingEnvironmentIdentifier recorded in the linking Object, Agent or Event (the linking relationships 1, 2, 3 and 5 depicted in Figure 1 pointing towards Environment). For the bi-directional relationships 2, 3, and 5 in Figure 1 one may use the linking[Entity]Identifier from within the Environment entity to identify related Objects, Agents or Events.

The question of whether Environment descriptions are stored as separate Information Packages in the repository or whether they must be stored together with the Objects or Agents whose role they take should not be specified within the PREMIS Data Dictionary since PREMIS is implementation independent. As with all implementations, however, if the PREMIS identifier mechanism is used, it must be guaranteed that it persistently and uniquely identifies the entity.

We are proposing a variety of mechanisms for implementing the relationship 4 depicted in Figure 1, which relates one Environment instance to another.

From within an Environment instance, one can refer to other Environments, such as from the description of a software application as Environment A to its operating system as Environment B. This would take the form of a relatedEnvironmentIdentifier

link using the PREMIS identifier mechanism to capture structural, derivative and dependency relationships.

Additionally, from within a local Environment instance in a repository one can refer to the corresponding (possibly, more complete or more up-to-date) descriptions in other registries (e.g. TOTEM or PRONOM). Here a premis:registryKey could be used to refer to information about the description in an external registry. Note that such a description does not imply identity between the Environment descriptions in the repository and the registry. Because of the sliding specificity of Environment descriptions (see Section 3.3) it is almost impossible to assert that two descriptions are identical. We assume that the referenced Environment description in the registry has to be more generic, and, therefore, can be inherited.

A further form of linking to an external Environment description could be an Environment designation, consisting of name and version. Additional specifications, such as the country of release of the version can be used to identify the Environment precisely.

In order to allow referring to different, internal or external descriptions of the same Environment at the same time, any form of linking should be repeatable and combinable. Each use of a linking mechanism should declare its role by some mechanism, such as premis:registryRole or linking[Entity]Role.

3.7 Expressing Dependencies between Environments

How Environments depend on each other so that they can be run, is key preservation information, which has to be expressed in the most satisfactory way possible. In PREMIS 2 dependencies can be expressed in two places:

1. DependencyIdentifier is used to document a non-software dependency between an Object and another Object, and uses an identifier mechanism to link to the required object.
2. swDependency expresses the fact that a piece of software, part of an Environment supporting an Object, relies on other software to be executed. This swDependency semantic unit is a “full text description” with no linking capability.

A gap analysis uncovered some areas for improvement. For example, low-level software Environments, like operating systems, rely on hardware to run. There is no explicit possibility in PREMIS 2 to document the nature of the dependencies. One can loosely record a hardware and software description in the same Environment container but not express the fact or the nature of their dependence. Links to repository descriptions are currently possible for file formats but not for other environment types. Specification of versions are possible for software, but not for hardware.

With the proposed PREMIS 3 change of Environment becoming a first-class PREMIS entity rather than a semantic container in the Object description, explicit linking mechanisms for describing dependencies can be used.

The existing ways of achieving the goal of expressing dependencies have to be simplified and re-factored so that they are as easy to use (for implementers) and to maintain (for the PREMIS Editorial Committee) as possible, while maintaining expressiveness.

PREMIS has a generic and powerful mechanism that allows linking two descriptions and assigning a type to the link. The two most generic semantic units are the linking[Entity]Identifier

and the relationship ones. They can both be used for linking Environments, maintaining the existing pattern that the former links two instances of different entities, and the latter links two instances of the same entity. Thus:

- Whenever there is the need to express the fact that a preserved Object or an Agent relies on an Environment to run, you use a linkingEnvironmentIdentifier mechanism;
- Whenever there is a dependency between two Environment instances, a premis:relationship with a new relationshipType of “dependency” can be used; this achieves the goal of the previous swDependency, and allows other dependencies, such as hardware dependencies, to be expressed as well. This is in addition to the structural and derivative relationships between Environments mentioned above. This implements the linking relationship 4 depicted in Figure 1.
- Whenever the dependency occurs between two Objects, the premis:relationship mechanism with the new relationshipType of “dependency” can be used between their Environments. This achieves the same purpose as the “dependencyIdentifier” PREMIS 2 feature described above.

The other advantage of this mechanism is its extensibility: the relationshipType and relationshipSubType semantic units’ recommended values in the Data Dictionary can be augmented. This is important as we cannot foresee all the relationships that can occur between Environments, which is a complex and evolving area. An example of a large variety of dependency relationships can be found in the Debian policy manual [28]. Using the relationship mechanism is a way to leave the door open to other relationships that could be needed in the future. Because of Environments’ highly interconnected, networked nature, the Data Dictionary solution should enable all of these linking and identification options.

3.8 Environments or Proxy Descriptions

When modelling Environments there is a decision to be made what form and content this Environment should take. If it will be preserved in an OAIS repository it will necessarily take the form of a digital bitstream, file or representation. Software and supporting documents, such as policy representations or manuals, can be captured directly in digital form as an Information Package. Hardware, business processes or non-digital documents are inherently not (necessarily) represented digitally and thus not directly subject to digital preservation as preservation Objects.

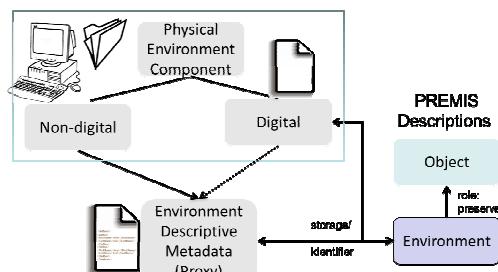


Figure 2: Environment components as preservation Objects

In either case, the object can be reduced to a proxy digital description that can be preserved as an Object. This descriptive environment metadata captures the physical object’s relevant characteristics and contains all the information needed to rede-

ploy a corresponding environment component with these same relevant characteristics in the future. This kind of environment preservation through proxy descriptions is used in, for example, business process preservation, as illustrated in the TIMBUS project [26]. See Figure 2 for an illustration.

A functional software description or the specification “Adobe Reader 5.0” can be considered instances of proxy descriptive Environment metadata. It is not as concrete as the Adobe Reader software composed of 0s and 1s, in the form of digital files that are the actual physical Environment component. Either or both could take the role of a premis:Object.

It is a business decision of the repository whether it preserves the actual digital representation of the Environment and/or Environment descriptive metadata as a proxy. This is a semantic issue. As with other curatorial decisions, this cannot be prescribed by the PREMIS data dictionary. But the eventual solution for PREMIS Environments must accommodate either use and allow for the nature of the Environment description to be specified.

3.9 Existing PREMIS Environment Descriptions

Keeping the existing solutions for describing Environments in PREMIS 2, the “environment” semantic unit and the “creatingApplication” semantic unit, enables backwards compatibility and, pragmatically speaking, offers convenient shortcuts and reduced verbosity for the situations in which they suffice. The PREMIS Environment working group does, however, feel that we would recommend the new Environment entity above those legacy semantic units.

4. USE CASE BASED DESIGN

The proposed solution is based on concrete examples rather than abstract considerations. It was driven by and validated with use case analysis. The working group validated that the modelling decisions, which were taken in extending the expressive capacities of PREMIS beyond the sheer description of preserved Objects to representation networks, were applicable to real-world examples.

Use cases should address all scenarios that implementers would expect to implement using PREMIS 3 Environments. The following examples were chosen:

- Describing the environment that is used to render web archives in a particular institution, with all the pieces of software that it bundles together to achieve this purpose;
- Describing the environment used in a normalization event;
- Describing the environment, including testbeds and documentation, used during TIFF to JPEG2000 migration;
- Describing an emulation environment for a Commodore 64 game preserved as an Object;
- Documenting the business processes in a multinational enterprise that operates in the cloud, and all the software and hardware dependencies that allow them to be re-deployed in the future.

The first two have been implemented in detail with a draft Data Dictionary proposal. With their help, it is possible to illustrate some of the features of the proposed Environment extension.

4.1 Use Case: Rendering Environments for Web Archives

In the first use case, harvested web pages from the web archives are rendered in the National Library of France’s reading room Environment. A web page harvested in 2010 can not necessarily be rendered on the reading room Environment of 2010. For example, for a web page harvested in 2010 that contains an EPUB file, this 2010 environment works for the HTML page. But the Firefox 2.0.0.15 browser it includes does not support EPUB files. The reading room Environment is upgraded in 2012 to an Environment that contains a newer version of Firefox that supports the EPUBReader plugin that allows one to render the EPUB file. In other terms, there was a need to describe these two Environments, the fact that one Environment is superseded by another, the different software components that they include, and the dependency relationships between them.

The preserved Object and its history are described with the PREMIS 2 standard features (Object, Event and Agent) as can be seen in Figure 3. The Environments are described separately and linked to from the Objects they support.

A new relationship type had to be introduced to state that the old Environment was superseded by the newer one. This information can, for example be used if the most current environment becomes obsolete. A preservation professional may choose to track superseded environments, which achieved the same purpose, in the hope of detecting a by-now readily available emulator of the older environment. This is an important feature for hardware and software preservation. This was achieved by a new relationshipType called “replacement”, with relationshipsubTypes of “supersedes” or “is superseded by”.

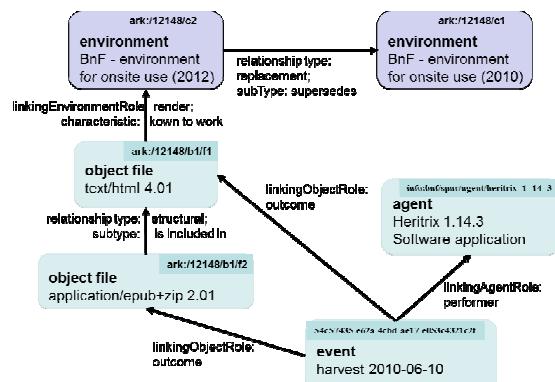


Figure 3: Web archive use case

This use case highlights how the environmentPurpose and environmentCharacteristics, familiar from the PREMIS 2 “environment” semantic unit, should be treated. The former was about the purpose an Environment wants to achieve towards a particular object (e.g. create, render, edit) and the latter, about the requirement that the Environment is intended to fulfil for a particular object (e.g. minimum service required, known to work). This should not be part of the Environment itself but part of the relationship between an Environment and the entity (Object or Agent) that it supports. This also increases the ability to share descriptions since the same Environment described above could potentially be used to achieve different purposes with different requirements.

Figure 4 shows the components of those two Environments. Each component is an individual Environment, and bundled into “aggregator” Environments. The aggregate mechanism, allows components to be shared across different Environments. For example, the Windows XP Service Pack 2 description is shared by both Environments since they use the same operating system.

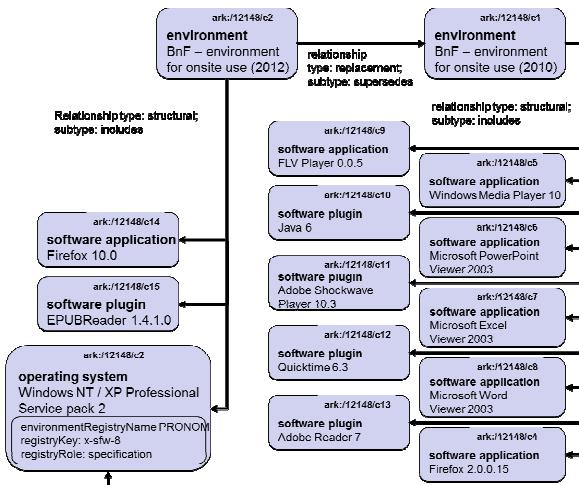


Figure 4: Inclusion links between Environment platforms and their component Environments

It also illustrates how one can link to a registry for additional descriptive information. Here, the Environment instance “ark:/12148/c2” describes Windows XP with a particular service pack; on the other hand, there is a description in PRONOM about Windows XP “in general”, with no particular service pack. In spite of this difference, adding this entry as a reference can be useful since the PRONOM description is likely to evolve and be enriched over time. A pointer to a repository should only be used if the description found there is an exact match or more generic and abstract than the Environment instance that links to it, so that the link does not cause conflicts in the Environment description.

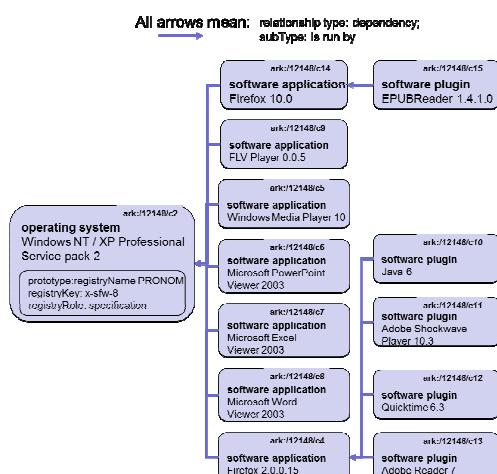


Figure 5: A dependency network between Environments

However, Figure 4 does not express all the required information. There is also the need to express the dependency relationships between the different components. Windows XP Pro-

fessional SP2, Firefox 10.0, and EPUBReader 1.4.1.0 are all part of the same aggregator Environment, but they do not act on the same level. EPUBReader, as an add-on, runs on Firefox 10.0, which in turn runs on Windows XP Professional SP2. These dependencies were documented by using another PREMIS relationship between the environments, as can be seen in Figure 5.

These two different relationships have to be distinguished because they do not act on the same level and do not achieve the same purpose. On the one hand, the whole/part structural links between Environments and their components are about picking Environment components to set up and bundle an Environment platform for a particular purpose, and are thus specific to a particular repository and implementation. On the other hand, the dependency relationships between the components are true whatever the context is.

4.2 Use Case: Documenting an Environment Used by a Normalization Service

In this use case, a QuickTime file with dv50 video and mp3 audio streams is submitted to a repository. Upon ingesting the QuickTime file, the archiving institution normalizes the file into a QuickTime file with mjepg video and lpcm audio streams. A normalization event is recorded, along with the web service and software that performed the format conversion. The derivation links between Objects, and their provenances are described by standard PREMIS entities and semantic units. The new feature is about the Agent description, which is a normalization service with no further description. So the Agent is linked to an Environment which describes what components are actually used by the service, e.g. libquicktime 1.1.5 with dependent plug-ins. The whole description can be summarized in Figure 6 below.

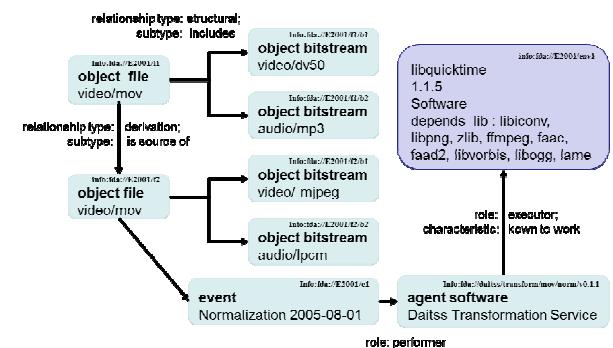


Figure 6: Normalisation use case

The distinction between the Agent and the Environment that executes it is important, if one wants to preserve an Agent so that it could be re-enacted in different Environments, or if one wants to track errors that have been discovered or link to an external registry. To this end, one may need to document the software components of which the Agent is built, along with the different Events that have been performed by this Agent in a repository. All this can be done by following the links between those different entities. This example also shows that different verbosity levels can be achieved depending on the implementer’s needs. While the web archives use case above used a very thorough Environment network description, this normalization example describes the execution Environment of an Agent more concisely. All the dependent libraries are listed in a single envi-

ronmentNote semantic unit. However it shall be noted that a more precise description could have been made if needed. In such a case, there would have been a distinct Environment description for each component (the software application and all its libraries), an inclusion link to an aggregator Environment executing the Agent, and, finally, dependency relationships between the libraries and the application. All depends on how far a PREMIS implementer needs, or wants, to describe Environments supporting the Objects s/he preserves or the Agents s/he uses. This ability to fit different needs is one of the key principles that guided this study.

5. CONCLUSION

The PREMIS Environment working group has been tasked with rethinking how a computing Environment should be modelled so that it meets the digital preservation community's requirements. Several open issues are still being investigated. The analysis and proposed solutions discussed in this paper will be brought to the PREMIS Editorial Committee and will be validated on community-provided use cases. Working within our stated modelling principles, we hope that our proposed approach not only meets contemporary registry preservation needs, but also improves the interoperability between Environment registries that are being developed within the community. The working group has included representatives from the PREMIS [3] Editorial committee, the TOTEM [21] technical registry, the IIPC [24], DAITSS [29] and the TIMBUS [26] project, and has received user requirements from New York University.

6. ACKNOWLEDGEMENTS

The authors wish to thank Michael Nolan, Martin Alexander Neumann, Priscilla Caplan and Joseph Pawletko for their contributions. Part of this work has been funded by the TIMBUS project, co-funded by the European Union under the 7th Framework Programme for research and technological development and demonstration activities (FP7/2007-2013) under grant agreement no. 269940. The authors are solely responsible for the content of this paper.

7. REFERENCES

Websites were accessed on 8 May 2012

- [1] Dappert, A., Enders, M. 2010. Digital Preservation Metadata Standards, NISO Information Standards Quarterly. June 2010. http://www.loc.gov/standards/premis/FE_Dappert_Enders_MetadataStds_isqv22no2.pdf
- [2] CCSDS, June 2012. Reference Model for an Open Archival Information System (OAIS): version 2. CCSDS 650.0-B-1, Blue Book (the full ISO standard). <http://public.ccsds.org/publications/archive/650x0m2.pdf>
- [3] PREMIS Editorial Committee, 2012. PREMIS Data Dictionary for Preservation Metadata, Version 2.2. <http://www.loc.gov/standards/premis/v2/premis-2-2.pdf>
- [4] Dappert, A., Farquhar, A. 2009. Modelling Organizational Preservation Goals to Guide Digital Preservation, Vol.4(2)(2009) of International Journal of Digital Curation. pp. 119-134 <http://www.ijdc.net/index.php/ijdc/article/viewFile/123/103>
- [5] Matthews, B., McIlwrath, B., Giaretta, D., Conway, E. 2008. The Significant Properties of Software: A Study. STFC, December 2008. <http://bit.ly/eF7yNv>
- [6] Software Sustainability Institute, Curtis+Cartwright. 2010. Preserving software resources. <http://software.ac.uk/resources/preserving-software-resources>
- [7] McDonough, J., et alii. 2010. Preserving Virtual Worlds. https://www.ideals.illinois.edu/bitstream/handle/2142/1709_7/PVW_FinalReport.pdf?sequence=2
- [8] POCOS. Preservation of Complex Objects Symposia. <http://www.pocos.org/>
- [9] SWOP. SWOP: The Software Ontology Project. <http://www.jisc.ac.uk/whatwedo/programmes/infl1/digpres/swop.aspx> and <http://sourceforge.net/projects/theswo/files/>
- [10] Dumbill, E. 2012. DOAP- Description of a Project. <http://trac.usefulinc.com/doap>
- [11] National Software Reference Library (NSRL). National Software Reference Library www.nsrl.nist.gov/
- [12] MobyGames. <http://www.mobygames.com/>
- [13] AMINET. <http://aminet.net/>
- [14] JSTOR and the Harvard University Library. JHOVE - JSTOR/Harvard Object Validation Environment. <http://hul.harvard.edu/jhove/>
- [15] The National Archives: PRONOM. <http://www.nationalarchives.gov.uk/pronom/>
- [16] Unified Digital Format Registry (UDFR) <http://www.udfr.org>
- [17] Library of Congress. www.digitalpreservation.gov/formats/
- [18] UDFR. UDFR ontology. <http://udfr.org/onto/onto.rdf>
- [19] Kadobayashi, Y. 2010. Toward Measurement and Analysis of Virtualized Infrastructure: Scaffolding from an Ontological Perspective. http://www.caida.org/workshops/wide-casfi/1004/slides/wide-casfi1004_ykadobayashi.pdf
- [20] SNIA. Cloud Data Management Interface (CDMI) <http://www.snia.org/cdmi>
- [21] W3C. 2001. Web Services Description Language (WSDL) 1.1. www.w3.org/TR/wsdl
- [22] KEEP (Keeping Emulation Environments Portable). <http://www.keep-project.eu/ezpub2/index.php>
- [23] TOTEM database. Welcome to TOTEM - the Trustworthy Online Technical Environment Metadata Database. <http://keep-totem.co.uk>
- [24] IIPC Preservation Working Group. <http://netpreserve.org/about/pwg.php>
- [25] TOSEC (The Old School Emulation Centre). What is TOSEC. <http://www.tosecdev.org/index.php/the-project>
- [26] TIMBUS project. <http://timbusproject.net>
- [27] Object Management Group. 2011. Business Process Model and Notation (BPMN). Version 2.0. Release date: January 2011. <http://www.omg.org/spec/BPMN/2.0/PDF>
- [28] Debian. Debian Policy Manual. Chapter 7 - Declaring relationships between packages. Version 3.9.3.1, 2012-03-04. <http://www.debian.org/doc/debian-policy/ch-relationships.html>
- [29] DAITSS. <http://daitss.fcla.edu/>

LDS³: Applying Digital Preservation Principles to Linked Data Systems

David Tarrant and Les Carr
School of Electronics and Computer Science
University of Southampton
Southampton
UK
SO17 1BJ
[davetaz.lac @ecs.soton.ac.uk](mailto:davetaz.lac@ecs.soton.ac.uk)

Data publishing using semantic web and linked data techniques enables the sharing of detailed information. Importantly this information is shared using common standards and vocabularies to enable simple re-use. In the digital preservation community, an increasing number of systems are adopting linked data techniques for sharing data, including the PRONOM and UDFR technical registries. In many systems, only current information is being shared. Further, this information is not being described with data relating to who and when it was published. Such basic metadata is seen as essential in all digital preservation systems, however has been overlooked to a large extent when publishing linked data. This failing is partly due to there being very few specifications, reference implementations and verification systems in place to aid with publishing this type of linked data. This publication introduces the Linked Data Simple Storage Specification, a solution that enables careful curation linked data by following a series of current best practise guidelines. Through construction of a reference implementation, this work introduces how historical information can be referenced and discovered in order to build customisable alerting services for risk management in preservation systems.

1. INTRODUCTION

Data, or to use another term, knowledge is the foundation for progression in society. Knowledge is key to making informed decisions that hopefully, on reflection, are correct. This principal is particularly true in the field of digital preservation and archiving where a key opportunity exists to automate the sharing of knowledge for the good of the entire community. The most common form of knowledge exchange within the digital preservation community is via registries ([7],[18],[8]). Moving on from simple fact based registries, such systems have evolved with the aim of sharing process information [1] to the point where it is now possible to share work-flows [10].

The automated sharing knowledge via the web is an area of research that has seen huge interest over the past decade, partly driven by the vision for a Semantic Web [4]. In this vision, knowledge comes together with reasoning such that informed decisions can be made on a persons behalf. This is a field of study which brings modern techniques together with years of Artificial Intelligence research [12].

The idea of publishing self describing data on the web, that could be read and understood by computers became the key driving principal for what is now known as Linked Data. Berners-Lee outlines a 5-star guide for publishing linked data on the web [3], a guide that has been followed successfully by many communities ([6],[13],[17]) including in the field of digital preservation [9].

The P2-Registry prototype [18] took advantage of the ability to harvest, manipulate and reason over linked data available from many sources to help make informed decisions regarding preservation actions. Data from PRONOM and DBpedia (the linked data version of wikipedia) was imported and aligned using a series of simple ontologies. This lead to huge increases in the amount of knowledge available to answer questions relating to specific digital preservation problems including: “What tools can open a particular file?”, and “How do I migrate this file to JP2000?”.

The original P2-Registry prototype has been utilised successfully by many preservation systems to help users make important decisions ([2],[19]). In addition many other linked-data related projects have began in the area of digital preservation, most notably the PRONOM data is now available directly from the National Archives (UK) as linked data [9].

While the amount of linked-data becoming available from various sources is becoming much greater, there still exists many problems in managing this data and deploying the correct architectures. Further challenges are then faced in understanding what information is available, establishing trust of this information and separating historical and current information.

While these problems exist within both the UK government data (where PRONOM is hosted) and P2-Registry system, they are not unique in these systems. In the years following the initial effort on the P2 system, many efforts have been made in the wider community to tackle the problems with understanding, trust and provenance resulting in the production of many best practise guidelines. In this publication, we present LDS³, the successor to P2 that follows a number of these best practices to provide a simple system which automates and assists with the process of publishing data to maintain integrity, trust and full historical information. Further to this, the LDS³ system also enforces strict

data curation policies, meaning any hosted datasets should be easy to understand, query and re-use.

LDS³ supports a publication-based named graph model to re-connect data indexed for querying to the actual source data. Further LDS³ removes the concern from the user about version and temporal data, much like version control systems do for computer code, enabling users to directly upload and manipulate documents containing the important data. The LDS³ reference implementation extends a number of freely available and well supported software libraries. This is done with a lightweight shim that simplifies and streamlines the process of managing linked data. At the same time as implementing the LDS³ specification, this shim also incorporates authentication services using OAuth2 to allow the management of data to be restricted.

This publication presents both the LDS³ specification and related reference implementation. Further a number of exemplar use cases, similar to that presented in the P2-Registry work, are introduced to demonstrate the benefits of the new capabilities available. Specifically, one of these capabilities looks at how historical information can be queried to provide automated alerting services when expected behavioural change.

The remainder of this paper is structured as follows. Section 2 recaps the P2-Registry and related work from the wider community, introducing many of the efforts being made to produce best practice guidelines for managing trust, authenticity and history of data on the web. Sections 3 and 4 introduce the LDS³ specification and reference implementations addressing how some of these best practise guidelines have been applied to produce a specification for managing data.

Section 5 looks at the problem with changing data in the digital preservation community. By continuing the P2-Registry work, this section looks at the risks to changing characterisation data and outlines how LDS³ can be used to build alerting services to information about risks related to change in this type of data. Before concluding the broader implications for LDS³ type systems are introduced demonstrating how LDS³ supports discovery and querying of historical data.

This paper concludes by looking at the applications of LDS³ and possible future work. This section looks at how the P2-Registry has now been enhanced with temporal data without changing the existing API and available services. LDS³ provides an exemplar for publishing persistent datasets that provide valuable information needed to establish trust,. By extending the use of such services beyond the preservation community, this will in turn enable easier data preservation in the future.

2. LINKED DATA TODAY

Berners-Lee's original vision for the Semantic Web became a vision for the future of automated computing in which information is not only discoverable and transferable, but also fully understood. Further, this information enables the generation of new knowledge through complex reasoning and other inferencing techniques. Essentially the web and http would be used as the location, storage and transport meth-

ods for knowledge. Artificial Intelligence methods would be required to assist with trust, proof and the understanding of the data.

While the semantic web is still a vision, some of the barriers to seamless knowledge exchange are being lowered. Sharing of knowledge starts with the sharing of data; facts that can be used in other contexts. The web has encouraged the sharing of information, however this has typically been via the embedding of data in web pages (using HTML). The drawback of this technique is that HTML is designed as a human readable format and not one to be used for automated exchange of understandable data. In order to move to a web of machine readable, open data requires a new way to expose data.

The benefits of sharing data have been seen in many applications [5]. Many services have opened up their data using formats such as XML, JSON and simple CSV, following the 5-star principals of linked data [3]. Exposing data under an open licence in this way achieves between two and three stars. The 4th star calls for the data to be shared in the RDF format, using URIs for identifiers, such that data can be easily discovered over the web and then used in a standards compliant way. Once the data is exposed as 4-star Linked Data, techniques from the Semantic Web can be used to align datasets from disparate sources, leading to a greater breadth of knowledge being available. 5-star Linked Data is that which is already aligned and linked in some way to other available 4 and 5-star linked datasets.

The idea of the P2-Registry was to expose the benefit of creating 5-star linked data for the digital preservation community. This was achieved through the linking of the PRONOM data to that exposed by DBpedia (the data endpoint for wikipedia). At the time the PRONOM data was not exposed as Linked Data, thus translating the XML data into RDF with URIs was necessary. This was required in order to get to a point where semantic web techniques could be used to align and link to the data from DBpedia.

Figure 1 shows the use of the RDF Schema vocabulary to connect two PRONOM identifiers (two versions of the PDF file format) to the DBpedia identifier for Portable Document Format. As DBpedia does not contain entries for each version of PDF, these links state that each PRONOM identifier is a subClass of the file format. In the case where a direct mapping could be found, i.e. for software URIs, then the sameAs predicate can be utilised from the Web Ontology Language (OWL) ontology.

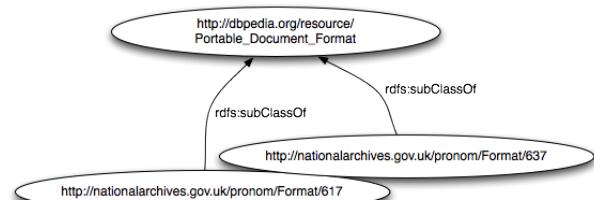


Figure 1: Associating PRONOM data with DBpedia data

The benefit of this simple link is easy to see when asking questions about the software tools available to read and write PDF files. With only the PRONOM data being used, the number of available tools was found to be 19. With the alignment to DBpedia (as shown in Figure 1), this number jumps to 70. Thus one connection (from PDF 1.4 to DBpedia) results in a near 4 fold increase in available data.

Since the P2-Registry work, the PRONOM data has been made available by The National Archives (UK) as 4-star linked data [9], with the 5th star (linking to other content) something of great interest. This work was enabled through the push in the UK for releasing of government data as 4 and 5-star linked data, something for which there is traction and now a substantial number of datasets available. International efforts have also been pushing to make raw data available in similar ways [8].

The publishing of linked data is just a single step towards fulfilling the promise of the semantic web. The problem is that the current methods for publishing and managing linked data fall short when looking at the full intention of the semantic web. Current publishing methods don't guarantee understanding, trust is not easy to establish and provenance information is also hard to find. Problems with establishing trust can be explained by analysing current publication and dissemination methods to discover that linked data is often only made available in a way disconnected from its source. When the source of the data is located, a process not made easy by current systems, it is still not clear how current and valid this data is, and what previous state the information held.

In the years following the initial effort on the P2 system, many efforts have been made in the community to tackle the problems with understanding, trust and provenance of linked data. This has resulted in the production of many best practise guidelines that are discussed in this section.

2.1 Publishing Linked Data

Publishing of linked data starts with knowledge modelling, the process of taking existing data and deciding how to serialise this into a linked data format, typically RDF. Take the following axiom of information:

```
<David_Tarrant> worksFor <University_of_Southampton>
```

While this is a valid triple, on its own no clue is given about the validity of this information, something normally established by looking at the information source (e.g. this publication). Once discovered, questions like "how old is this information?" and "who published this information", can be answered easily. However in linked data (using RDF or SPARQL), it is not clear how to find the source of such information.

This was realised as problem by early linked data systems, examples of which include triple-stores. Such systems would store a fourth piece of information detailing the location from which the information originated so it could be easily updated. While systems designed to index and store linked data realised this need, it is still not fully realised by systems that expose this data, as was the case in the P2-Registry.

Many active linked-data systems utilise storage and indexing systems as their only dissemination mechanism, often with an accompanying SPARQL (RDF Query Language) endpoint. While this allows the data to be re-sliced to answer queries, this results in a disconnection between the exposed data and the original sources. In the P2-Registry, answers to queries consisted of data from two data sources (PRONOM and DBpedia), resulting in this same disconnection problem.

Moving from a triple based RDF model to that of a quad, means that named graphs (term for the quad), can be used to provide source information. Named Graphs can be used in two ways, either to express publication information or for representation information [17]. Using named graphs to express publication information allows the connection back to the original source (here termed as publication). Representation information relates more directly to the result of combining data, e.g. the source of a query and data about the query endpoint. There is value in both uses, especially as it may be required to keep a record of where the data was discovered (or queried from) as well as the locations for the original sources of that data.

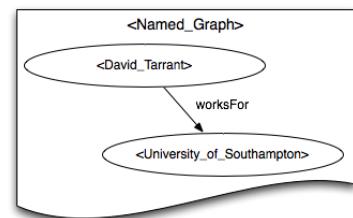


Figure 2: Encoding a triple with a named graph (a quad)

Figure 2 shows an example of the previous triple now represented with a quad. In the case of this representation it has been chosen to represent the named graph as a document that represents the source of the triple. Equally this document might convey information relating to many subjects (in this case people) and their related information.

Taking this forward, Figure 2 also indicates that the `<Named_Graph>` can also be the subject of information, thus allowing triples to be included in this named graph that describe itself. It is this data that can include facts like the author, publisher and publication time.

Exposing the named graph in queries immediately allows separation of data sources, allowing data from PRONOM to be differentiated from that produced via wikipedia. Knowing the exact source of the data allows any user to retrieve the original data from its source (rather than the query endpoint) in order to verify the information and establish some level of trust. Additionally, techniques such as Public Key Identifiers (PKI) can also be used at this point to further verify that the data received is authentic [14].

Using named graphs for publication data clearly has its benefits, but requires that a user be able to retrieve the original data for inspection, not via an index of the data. While

sources for information, e.g. RDF, can be easily hosted on web servers directly, this process relies on the user to keep these documents up to date and properly annotating them with information relating to the time and place of publication. As well as combining indexing and query services, the main role of LDS³ (as a Simple Storage Service) is to provide hosting services for the source of data. LDS³ enforces the use of named graphs to represent publication data and will automatically annotate data that it is hosting with the publisher and publication time data, meaning that the author does not have to worry about these aspects. Providing both storage and indexing services means that LDS³ is able to easily keep the two services in synchronisation while allowing users easy access to the source documents that were used to build the index.

2.2 Versioning Linked Data

The publication of linked data is typically a two-stage process involving the initial creation and subsequent importing of data into a linked data endpoint. It is this endpoint that provides fast access to the latest version of information using direct export or query functionality [17]. Further, such endpoints all include functionality for managing data indexes and ability to apply simplistic semantic reasoning. These systems were the early adopters of named graphs, using this information to allow data to be updated and overwritten, allowing the index to only return the most up to date (and thus valid) results. This is perfectly acceptable as the majority of queries are asking for current data. With many systems regarding the data endpoint as the only way to access data, finding previous information can be a significant challenge.

The problem with versioning resources is not necessarily applicable to all resources, for example statistical data intrinsically relies on temporal and contextual data to justify its own results. On the semantic web, such data would be referred to as an information resource. On the other hand data about a University, or Person, is an example of a non-information resource, where the main requirement is to discover current information. [17] (also discussed on Jeni Tennison's blog¹) examines the problem with versioning information and non-information resources. One of the main conclusions is that it should be possible (not necessarily easy) to discover the previous state of non-information resources.

One technique for versioning linked data relating to non-information resources is to use publication named graphs these. Tennison recommends combining named graphs with cool URIs [16], making it very easy to see that versioning is being used. Further these URIs can be used to relate versions together, as it demonstrated in the example below:

```
<http://data.ac.uk/doc/{resource}/{version-2}>
  dct:replaces <http://data.ac.uk/doc/{resource}/{version-1}>
  dct:published '2012-05-09 14:00:00+01:00'
  dct:author <http://id.ecs.soton.ac.uk/person/9455>

<http://id.southampton.ac.uk>
  foaf:Name 'University of Southampton'
```

Here the resource name and versioning scheme can be freely defined by the publisher, such that schemes such as

¹Versioning (UK Government) Linked Data - <http://www.jenitennison.com/blog/node/141>

simple version numbers can be used, or perhaps the date of publication is embedded in the named graph URI. Importantly, by using already available technologies, it is possible to navigate easily between versions of a named graph that (potentially) contain information relating to an Information Resource published by the same author, akin to editions of a book.

By separating storage from indexing, LDS³ automatically creates and manages versions of named graphs submitted by authors. This way all previous versions of a named graph, containing all original data are available from storage, with the latest version available directly from the index. LDS³ adopts a combination of Globally Unique Identifiers (GUID) and date stamps to generate the named graph URIs and versions of this URIs respectively. This also allows a user to ask for a GUID (without a date) and be re-directed automatically to the latest version.

In the field of digital preservation, people have for many years been talking about registries as the source for information. However these registries contain the same flaws due to the lack of temporal and provenance information. Historically (before digital), a register is a book in which records are kept, thus the authoritative source of information may well be a page in this book and cited in the same way as traditional journals. Each register would have its own version information and publication date. As registers have become digital, it has become very easy to duplicate and move data around and simply overwrite old data, losing the versioning and authoritative information related to the original publisher. In part this is due to the lack of clarity on what is the source of data, and what is simply a representation built from some index (or registry). Using named graphs effectively re-introduces versioned registries, where a much greater level of granularity is possible.

3. THE LDS³ SPECIFICATION

The Linked Data Simple Storage Specification² outlines a mechanism for assisted publication of linked data. By taking influences from many existing systems, LDS³ and accompanying reference implementation enables the management and exposure of large scale datasets. The LDS³ specification utilises the named graph as a publication reference and requires any compliant server to automatically augment incoming data with further information relating to both the time and author responsible for the publication. All requests to publish data must be authenticated in a secure manner before data is augmented and URIs returned to the requestor.

The LDS³ specification takes many influences from existing specifications, most notably the AtomPub[11] and SWORD2³ specifications. These existing specifications focus on the publishing of web and scholarly resources respectively. LDS³ compliments these specifications while focussing on data publication and providing services to help with the curation and automated tracking of versions.

²LDS³ Specification - <http://www.lds3.org/Specification>

³SWORD2 (Simple Web-service Offering Repository Deposit) Specification - <http://swordapp.org/sword-v2/sword-v2-specifications/>

The most important influence from both the AtomPub and SWORD2 specifications is the reference to CRUD (Create, Retrieve, Update and Delete) for managing resources. Each create request will also be processed to generate specific objects (and related URIs) within the LDS³ system.

The process of creating a resource is shown by Figure 3 where data is HTTP POST'ed to the servers Data Submission endpoint. The server handles the request in the standard HTTP based way and simply returned the location of the created resource. Further to this location, the server also returns the *edit-iri* that can be used to update and delete the document.

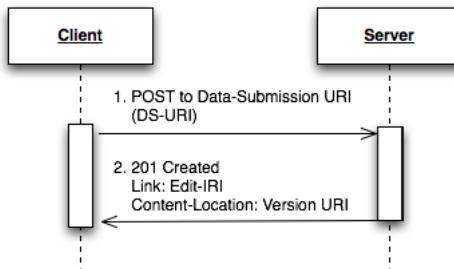


Figure 3: Submitting a new named graph to LDS³

For authentication, LDS³ requires that all requests be signed using the same technique as employed by Amazon's Simple Storage Service (S3)⁴. This key based authentication mechanism works by users signing parts of the HTTP request with their private key. With only the request part of the transaction being signed, the process of authentication does not require bi-directional communication, meaning no loss in performance.

3.1 Managing resources with LDS³

An implementation of LDS³ is intended to be deployed directly on the web server hosting the data URIs (e.g. starting `id.data.ac.uk`). This way the LDS³ implementation can directly serve requests for information and non-information resources as well as the named graphs. Information relating to resources is likely to be sourced from many documents, thus requests for a resource will be handled via the index of the latest data. Named graphs, both current and previous versions can be provided directly from disk, avoiding the need for a data index.

While specifications for handling data indexes are well defined, LDS³ compliments existing systems by also handling the publication of the named graphs, annotating these and storing them for indexing and provisioning to other systems and users. The LDS³ specification dictates that resources (e.g. People, Universities or File Formats) cannot be directly created, updated or deleted. Each resource has to be described in a published document (named graph). This paradigm is similar to that of traditional publishing, where the trust of information is to some degree established by looking at the Book, Journal or Proceedings in which the

⁴Signing and Authenticating REST Requests - <http://docs.amazonwebservices.com/AmazonS3/latest/dev/RESTAuthentication.html>

data was published. By limiting users to only being able to publish and update documents, the LDS³ enforces a model of versioning and provenance on resources. These graphs are thus being used as the publication mechanism rather than as presenting representational information.

Figure 2 shows how one document can be used to describe a resource. Here the LDS³ endpoint is hosting data at `http://data.opf.org/`, with non-information resources having a prefix of `http://data.opf.org/id/`. Note that in Figure 2 the named graph URI is an example URI. When an LDS³ server receives a correctly formatted an authenticated request, a unique URI must be created for the document. This URI should consist of two parts, one to identify the document series (the aforementioned *edit-iri*), the other for the version of the document. It is recommended to use a GUID for the *edit-iri* and append a version or date to this URI as the location of this particular version of the document.

Taking Figure 2 from before, the server then fills in (or changes) the document URI to the new URI and annotates it with data pertaining to who published the document, when and which (if any) documents it replaces. This results in a new document being generated similar to that shown by Figure 4.

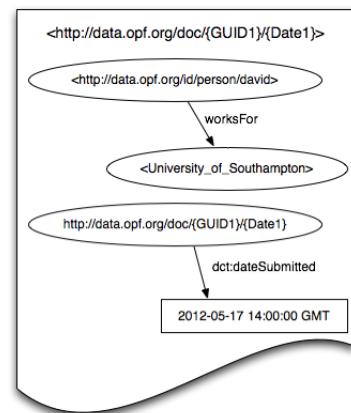


Figure 4: Document including LDS³ annotation

Figure 4 shows that the named graphs are stored under their own document prefix `http://data.opf.org/doc/` with GUID and Data used as the suffix's (not shown here in order to save space). As well as the submitted data from Figure 2, the LDS³ system has annotated the graph to make it self describing, adding the date when the graph was submitted. It is this exact same mechanism that is used to add further annotations and links to previous versions of the document.

Once the data has been annotated and stored, the *edit-iri* (or GUID only URI) is returned to the user along with a final representation of the annotated document. The final representation of the document is accompanied with the HTTP Content-Location header which defines the exact location on the server of the document, in this case the full document URI including GUID and Data suffixes. The *edit-iri* is communicated to the user using an HTTP Link header

(as shown in Figure 3), it is this URI that can be used to submit new versions of the document as well as retrieve the latest version.

Using named graphs in this way allows many users to submit data relating to the same URI whilst retaining the separation of who submitted what information. Correspondingly, as users can only manipulate documents, they are only able to delete the data that they added, and not directly manipulate the resource URI. It is a combination of these factors that mean LDS³ is able to provide enough information to enable the establishment of trust in the data. Allowing users to annotate their own named graphs and by enforcing versioning, allows the easy discovery of provenance information.

4. REFERENCE IMPLEMENTATION

In order to aid the deployment of LDS³, a reference implementation has been developed. Rather than start from scratch the reference implementation ties together many existing libraries. The only new piece of development involved the creating of a shim to handle authentication, requested operations and document annotation.

The authentication module requires that users register in order to obtain a key-pair. It is expected that this key-pair be used by the user's client in order to upload a series of documents, much like handling of objects in Amazon S3. Each key-pair remains linked to a single user account, but each user can have several key-pairs. To avoid building a user management system, the LDS³ reference implementation contains an OAuth2 [15] module, allowing any OAuth2 compatible authentication service to be used.

Once a user has a key-pair, documents can be submitted to the Data Submission IRI (DS-IRI). Each received request is verified before a new GUID is generated and added to the document. Annotation is performed using this Graphite library⁵ before storing the resultant document on disk and calling the index process to update the query endpoint. To index and allow querying of the data, the reference implementation recommends use of a quad store (such as 4store). Currently the LDS³ reference implementation only indexes the latest data, handling old versions of is discussed in section 6.

With the index in place and data ingested, the major requirement is to expose the datasets and make available a version of the Linked Data API⁶ to make the data usable. In order to achieve this, the Puelia-PHP library has been chosen and themed with the data.gov.uk style. data.gov.uk utilises the exact same set of libraries as LDS³, thus streamlining the functionality and mechanisms for publishing datasets, something also handled using Puelia-PHP.

Puelia-PHP is an application that handles incoming requests by reading a dataset configuration file to discover how to serve the request. Each dataset configuration outlines the URI pattern to match and how to query for the data from a SPARQL endpoint. The advantage with this type of de-

⁵Graphite - <http://graphite.ecs.soton.ac.uk/>

⁶The Linked Data API -<http://code.google.com/p/linked-data-api/>

ployment is that Puelia-PHP can gather different datasets from many SPARQL endpoints, spreading the hardware and processing requirements for hosting billions of items of data. Further the data is then cached to enable fast delivery for future requests. Finally, Puelia-PHP provides multiple serialisations of the data including JSON, XML, CSV alongside HTML and RDF.

As Puelia-PHP is designed to query data from a SPARQL endpoint and then serialise this into a new representation, the ability to retrieve the original named graph is not available. To counteract this, the LDS³ reference implementation recommends that Puelia-PHP be patched to enable retrieval of named graphs from either the precise document URL (.rdf), dated URI or related edit-IRI (both content negotiated). Since resource URIs cannot be directly edited, the use of a representational named graph here is ideal.

Although Puelia-PHP does provide an excellent and well supported implementation of the linked-data API, it currently lacks the ability to expose named graph information. This is due to the challenges in exposing non-native named graphs that are linked to non-information resources. The linked data API specifies that systems should be able to query many indexes to location information from many sources and aggregate this into a new named graph (a representation named graph). Options exist to simply use this new named graph to point to all the existing named graphs, resulting in a meta-aggregation that doesn't directly describe the object the user asked about. Further you can envisage infinite meta-aggregations, making the process of retrieving any piece of information a painful one.

```
SELECT * WHERE {  
Graph ?graph { ?subject ?predicate ?object }  
}
```

While SPARQL supports the direct retrieval of named graph information (as shown by the query above) it is the serialising of this information, into formats including RDF, that is the challenge. Not being able to serialise the data back to RDF doesn't mean that it cannot be used however and many other visualisation tools, including DISCO⁷ and MARBLES⁸ enable the browsing of quad based information. It is hoped that in the near future that this level of browsing capability can be bought to Puelia-PHP.

5. LEARNING FROM THE PAST

Historical records consist of two important pieces of information: facts about the environment at the time and decision data about choices made based upon interpretation of these facts. Example facts might include file format identification information (at the time), while process information outlines the actions, or provenance data, related to how these facts was used. It is the facts that inform the process, neither piece of data is useful without the other. Another way to look at facts, is to refer to them as non-information resources, while your processes are examples of information resources. Non-information resources (facts) can change over time, so only keeping the latest information means that the

⁷Disco Hyperdata Browser - <http://www4.wiwiiss.fu-berlin.de/bizer/ng4j/disco/>

⁸Marbles - <http://marbles.sourceforge.net>

information resources (processes) become a lot less useful.

A good example of non-information resources in the field of digital preservation is identification data. Many file format identification tools exist, each under continuous development as new formats and format types become available. Due to the dynamic nature of file formats, there is a high risk of miss-identification. This is particularly true with formats which re-use the zip and xml standards for packaging. Additionally there is a chance that older formats may get re-classified if a newer format is very similar. These are all high preservation risks, and ones that require services to inform people of change.

As part of the European project looking at Scalable Preservation Environments (SCAPE), an LDS³ implementation is being set up to store results of running a number of identification tools over a wide ranging corpora of exemplar data. By collecting this data over time, it will be possible to observe the changing behaviour of the tools and any potential risks to the identification process each version of a tool might introduce. For example, a number of the DROID signature files wrongly identify the Microsoft Word docx format, while other miss-identify PDF. Such information is currently only available to those running their own experiments, or via a few forums and mailing lists. There is currently no method for auto discovery of this information. By using LDS³ to store data relating to these experiments, it is possible to discover these risks and report on them automatically using Preservation Watch services (also being developed within the SCAPE project).

By gathering results from experiments, data from many sources, and combining this with temporal data. LDS³ has the capability to enhance the previous risk analysis work by being able to present evidence relating to how results have changed over time. Figure 5 shows the components of the preservation watch service for characterisation change, with an LDS³ system collecting the results ready for analysis and publication.

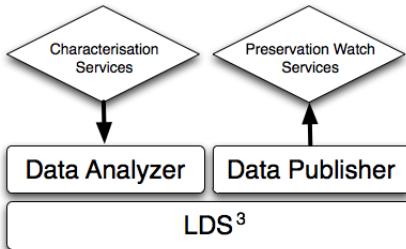


Figure 5: LDS³ and preservation watch services

This system involves many of the components being developed as part of the SCAPE project being developed by many different parties. LDS³ plays an important role in being a persistent store for published and usable datasets. By using widely available standards and technologies means that the many different parts of the system can be worked on independently to produce a usable solution for preservation practitioners.

When complete it is envisaged that the preservation watch services will produce a series of customisable alerts tailored for each individual user. If the users interest is in preserving multimedia content, then received alerts can be customised to only be relevant to this type of material. Most importantly though, each user will have the ability to trace the complete provenance of each alert, including the decision process and the facts that informed this alert. Further this can be done at any point in time, thus decisions made today, can be analysed again in the future without loss of information.

6. THE DATA TIME-MACHINE

The real appeal from this provenance information comes from what can be done with it, firstly and most obviously the clock can easily be turned back to discover the previous state of any named graph. As demonstrated by Figure 6, this can then be combined with the user interface to create a clear view of the data against time.

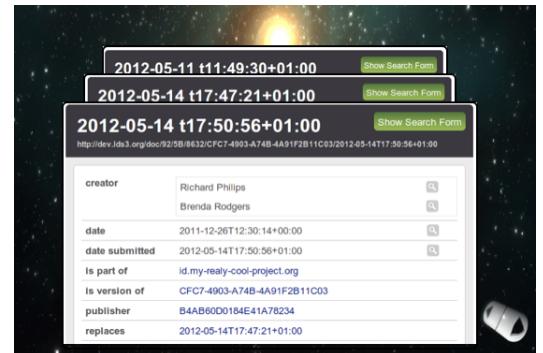


Figure 6: “Time-Machine” interface for LDS³

This “Time Machine” style interface for linked data, shown by figure 6 working with LDS³, allows the retrieval of any named graph from any point in time. This can also be achieved by using the Memento API [20] directly on the LDS³ server. The request below shows an example request for a document (e.g. that from Figure 4) at a specific point in time. Note that the only extension to the normal HTTP request is the addition of the “Accept-Datetime” header as defined by the Memento specification [21].

```

GET /doc/GUID1 HTTP/1.1
Date: Mon, 14 May 2012 15:55:15 GMT
Accept: application/rdf
\textbf{Accept-Datetime: Thu, 21 Jan 2012 04:00:00 GMT}
Host: data.example.org
  
```

In addition to providing access to static documents from the past, by maintaining a few indexes of named graphs and their relation to resources it is possible to rebuild an index as it looked at any point in time. This allows full SPARQL queries to be executed on the data as it existed at this point. This capability represents a breakthrough for retrieving the previous state of a resource. All current web archives are very static in nature, showing content conforming to how the harvesting service retrieved it. Being able to completely re-query the index as it looked at a specific time is a major improvement on this technology.

7. CONCLUSION AND FUTURE WORK

Exposing linked data specifically about digital preservation has already been shown to have large benefits for the community. The P2-Registry work demonstrated how a simple relation between two existing datasets results in a four fold increase in results for a query. However these results are always considered current and come without any provenance information relating to the origin of each individual result. A four fold increase in data is only possible if many people can describe the same (or similar) objects on the web and without provenance information it is impossible to establish trust in such distributed data. Additionally, if decision processes are made based upon this data, without access to historical information, it is challenging to review such decisions again in the future.

By focussing on identification information, this publication presented a scenario in which the historical nature of identification information is not known. A major problem if a file format is wrongly identified. Such a change could cause serious consequences if process information is affected and called into question.

In order to address the challenges of provenance, versioning and trust, this publication introduced the Linked Data Simple Storage Specification (LDS³) and related reference implementation. LDS³ enforces the use of named graphs for publication of data related resources on the web, e.g. file format data. It is these named graphs that can be directly annotated with additional data including author, publisher and date of publication. Further, by using a combination of Globally Unique Identifiers (GUIDs) and time stamps in the URI scheme, LDS³ provides automatic versioning of data.

LDS³ provides an HTTP CRUD based interface enabling the secure management of fully annotated and versioned linked data. The LDS³ reference implementation, written as a shim, uses many existing and well supported libraries to perform data management, annotation and indexing. One such library, Puelia-PHP (used by the UK Government open data project), is used as the primary user interface with a quad-store backing the SPARQL endpoint.

As well as an LDS³ endpoint being created to store results of identification experiments, enabling the provisioning of preservation watch services, the existing P2-Registry system will be upgraded. This will enable sources of data to be discovered, allowing users to separate wikipedia data from that delivered by PRONOM. As the P2-Registry system was also based on the linked data principals, the user facing functionality and API does not change, it simply gets upgraded with new functionality designed to enable the establishing of trust and validity of data.

Having ingested fully annotated and versioned data. The LDS³ reference implementation applies parts of the Memento protocol to enable resources to be retrieved as they existed at specific points in time. Additionally it was demonstrated how these can be displayed in a “Time Machine” like user interface. Future work will investigate the possibility of allowing SPARQL queries to be performed over whole datasets, enabling fully dynamic query of semantically annotated datasets at any point in their history.

8. REFERENCES

- [1] B. Aitken, P. Helwig, et al. The planets testbed: Science for digital preservation. *Code4Lib Journal*, 2008.
- [2] C. Becker, H. Kulovits, et al. Plato: a service oriented decision support system for preservation planning. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, 2008.
- [3] T. Berners-Lee. Linked data. *w3c Design Issues*, 2006.
- [4] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 2001.
- [5] C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.
- [6] C. Bizer, J. Lehmann, et al. Dbpedia-a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2009.
- [7] A. Brown. Automating preservation: New developments in the pronom service. *RLG DigiNews*, 2005.
- [8] U. Center. Unified digital format registry (udfr). 2012.
- [9] R. Fisher. Linked data pronom. *National Archives Labs*, 2011.
- [10] C. Goble and D. De Roure. myexperiment: Social networking for workflow-using e-scientists. In *Proceedings of the 2nd workshop on Workflows in support of large-scale science*, 2007.
- [11] P. Hoffman and T. Bray. Atom publishing format and protocol (atompub). *IETF, RFC 5023*, 2006.
- [12] I. Horrocks, P. Patel-Schneider, and F. Van Harmelen. From shiq and rdf to owl: The making of a web ontology language. *Web semantics: science, services and agents on the World Wide Web*, 2003.
- [13] G. Kobilarov, T. Scott, et al. Media meets semantic web: How the bbc uses dbpedia and linked data to make connections. *The Semantic Web: Research and Applications*, 2009.
- [14] E. Rajabi, M. Kahani, et al. Trustworthiness of linked data using pki. In *World Wide Web Conference (www2012)*, 2012.
- [15] D. Recordon and D. Hardt. The oauth 2.0 authorization framework. *IETF*, 2011.
- [16] L. Sauermann, R. Cyganiak, and M. Völkel. Cool uris for the semantic web. *W3C Interest Group Note*, 3rd Decemeber 2008.
- [17] J. Sheridan and J. Tennison. Linking uk government data. *Statistics*, 2010.
- [18] D. Tarrant, S. Hitchcock, and L. Carr. Where the semantic web and web 2.0 meet format risk management: P2 registry. *International Journal of Digital Curation*, 2011.
- [19] D. Tarrant, S. Hitchcock, et al. Connecting preservation planning and plato with digital repository interfaces. In *7th International Conference on Preservation of Digital Objects (iPRES2010)*, 2010.
- [20] H. Van de Sompel, M. Nelson, et al. Memento: Time travel for the web. 2009.
- [21] H. Van de Sompel, M. Nelson, and R. Sanderson. Http framework for time-based access to resource states: Memento. *Internet Engineering Task Force*, 2010.

An Architectural Overview of the SCAPE Preservation Platform

Rainer Schmidt
AIT Austrian Institute of Technology
Donau-City-Strasse 1, Vienna, Austria
firstname.lastname@ait.ac.at

ABSTRACT

Cloud and data-intensive computing technologies have introduced novel methods to develop virtualized and scalable applications. The SCAPE Preservation Platform is an environment that leverages cloud computing in order to overcome scalability limitations in the context of digital preservation. In this paper, we provide an overview of the platform architecture and its system requirements. Furthermore, we present a flexible deployment model that can be used to dynamically reconfigure the system and provide initial insights on employing an open-source cloud platform for its realization.

1. INTRODUCTION

The SCAPE project is developing tools and services for the efficient planning and application of preservation strategies for large-scale, heterogeneous collections of complex digital objects [4]. The SCAPE Preservation Platform, developed in this context, provides the underlying hardware and software infrastructure that supports scalable preservation in terms of computation and storage. The system is designed to enhance the scalability of storage capacity and computational throughput of digital object management systems based on varying the number of computer nodes available in the system. It supports interaction with various information and data sources and sinks, the coordinated and parallel execution of preservation tools and workflows, and the reliable storage of voluminous data objects and records. At its core, the SCAPE Platform functions as a data center service that provides a scalable execution and storage backend which can be attached to different object management systems using standardized interfaces. The architecture aims at addressing scalability limitations regarding the number and size of the managed information objects and associated content.

The SCAPE preservation platform also supports a flexible software deployment model allowing users to reconfigure the system on demand. Packaging, virtualization, and automated deployment of tools and environments plays an im-

portant role in this context. A SCAPE preservation workflow may depend on dozens of underlying software libraries and tools, which must be made available on the computing infrastructure provided by the Platform. This in turn drives the need for a strategy that can resolve such context dependencies in a distributed and dynamically scaling environment on demand without requiring to perform expensive data staging operations over a network. SCAPE is employing a consistent packaging model in order to manage and sustain the preservation components developed within the project. Packaging and virtualization provide important concepts for the deployment and operation of the SCAPE Preservation Platform (and the tools and environments it depends on). In this paper, we describe a fully virtualized prototype instance of the SCAPE Execution Platform that has been recently set-up at AIT. Using a *private cloud* model, it allows us to deploy required preservation tools together with a parallel execution environment on demand and co-located with the data.

The rest of the paper is organized as follows: we provide an overview of the platform architecture and its key concepts in section 2. Section 3 presents design considerations for the packaging and deployment model. Section 4 discusses the application of data-intensive computing frameworks. A prototype setup of the preservation platform is presented in section 5. Section 6 reviews related work and section 7 concludes the paper.

2. ARCHITECTURAL OVERVIEW

The SCAPE Preservation Platform provides an digital object management and computing platform that (a) interacts with other SCAPE sub-systems like the Planning and Watch and the Result Evaluation Framework, and (b) supports the efficient execution and coordination of SCAPE *action components* like preservation tools and workflows.

2.1 Main System Entities

The main entities that make up the SCAPE Platform are the Execution Platform and the Digital Object Repository.

2.1.1 Execution Platform

The SCAPE Execution Platform provides a tightly coupled data storage and processing network (called a cluster) that forms the underlying infrastructure for performing data-intensive computations on the SCAPE Platform. The Execution Platform specifically supports the deployment, identification, and parallel execution of SCAPE tools

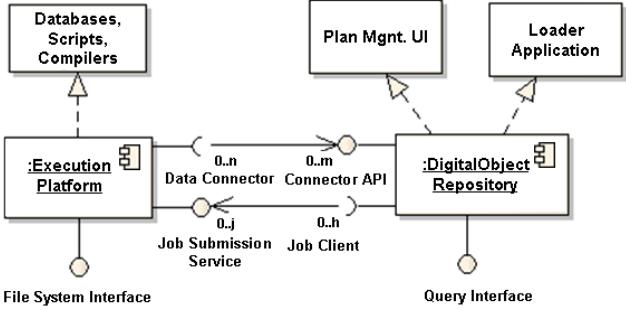


Figure 1: The SCAPE Platform comprises of two distinct system entities. The Execution Platform provides system-level support for storage and execution. The Digital Object Repository provides user-level support for data management and active preservation.

and workflows, and integrates with different data sources and data sinks. The system provides a set of command-line tools that support users in directly interacting with the system, for example to carry out data-management and preservation actions on the cluster. The Execution Platform does not provide graphical user interfaces per se but provides the below-described services to interact with client applications.

2.1.2 Digital Object Repository

A SCAPE Digital Object Repository (DOR) provides a data management system that interacts with the Execution Platform for carrying out preservation actions. The SCAPE DOR exposes also services to other entities developed in the context of SCAPE, for example for the Planning and Watch components. A SCAPE Digital Object Repository exchanges information with the Execution Platform via a defined API and may store or replicate its content directly to the Execution Platform's storage system. The DOR understands and manages Preservation Plans [2], triggers their execution, and may report to the Watch component. The repository may choose to preserve portions or the entire outcome of a workflow that has been executed against the content a DOR manages. It is therefore required that a DOR employs a corresponding data model as well as a scalable object store. Moreover, the object repository is responsible for aiding its user community in depositing, curating, and preserving digital content. A SCAPE repository reference implementation that is integrated with the platform's storage and execution environment is presently under development [1].

2.2 Interfaces and Services

Although the SCAPE Platform is designed to support software components and services provided by other SCAPE Sub-projects, its core entities may operate also independently from external services. A particular preservation scenario, for example, can be carried out autonomously once all required prerequisites (like data, tools, workflows) have been made available on a Platform instance. Figure 1 shows the interdependencies between the object repository and execution platform of the SCAPE Preservation Platform. The entities may interoperate with another using two defined ser-

vices; (1) the data connector API, and (2) the job submission service. These services represent the two core functionalities of the SCAPE Platform: data management and computation. Although a typical Platform deployment might involve only a single repository and a single execution platform, the system is not limited to this configuration. Both, the Data Connector API and the Job Submission Service maintain an n:m relationship with their clients.

2.2.1 Data Connector API

The Data Connector API provided by the DOR is a service that allows clients to efficiently create, retrieve and update digital objects. The interface is specifically designed to support bulk data exchange allowing clients for example to access data directly through the storage system. The connector API is used by the Job Execution Service to efficiently obtain and update content and metadata from the repository that manages a particular information object. The execution platform resolves data based on references and may access data from different repositories or other data sources. Additionally, one repository can supply data to multiple (perhaps differently configured) clusters. A job execution can also be performed independently from a DOR and only rely on the Platform's internal data management component (e.g. a distributed file system and/or database).

2.2.2 Job Execution Service

This service provides an interface for performing and monitoring parallel data processing operations (jobs) on the platform infrastructure. The object repository acts as a client to this service in order to actively perform preservation operations (as for example defined by a preservation plans) against the data it manages. Depending on the repository implementation and use-case, the processed data may or may not reside on the Platform's storage network prior to the execution. A job execution service can be utilized by multiple clients and/or repositories. Also, a digital object repository may use the Platform's storage network without implementing a client to the job execution service. On the other hand, a SCAPE Digital Object Repository may maintain its own storage layer and use the execution platform only on demand. An example for a loosely integrated repository is the implementation of an active cache that can be used for performing scalable preservation activities like for example file identification without directly exposing the storage layer of the repository.

3. INFRASTRUCTURE DEPLOYMENT

The SCAPE Platform architecture does not prescribe a specific deployment or infrastructure provisioning model. The system may be set-up using a private or institutionally shared hardware infrastructure, or be hosted by an external data center. The architecture can also take advantage of virtualization and can be deployed on a private or public IaaS infrastructure. Depending on its level of integration, a single Platform instance may also be shared between multiple tenants. In SCAPE, a central deployment of the preservation platform, called the SCAPE Central Instance, provides a secure and project-wide shared hardware and software environment for evaluation and demonstration purpose. At the time writing this paper, a number of private instances of the SCAPE platform are being set-up at institutions participating the SCAPE project.

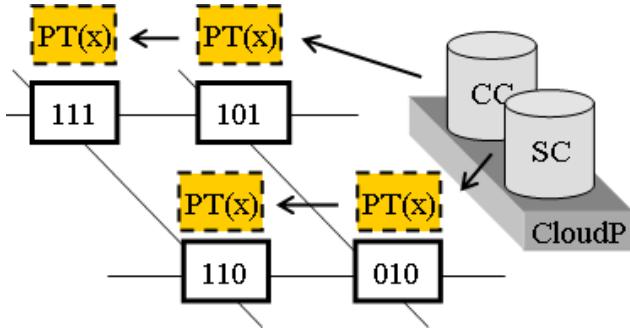


Figure 2: A setup for hosting the SCAPE Preservation Platform as a private cloud environment. Data is stored on persistent storage partitions directly on the nodes. The SCAPE software environment (PT) is deployed using transient virtual machine instances on demand on the nodes using a cloud platform.

3.1 Packaging Complex Environments

In addition to packaging software components, SCAPE is making use of virtual machine images in order to package complex software environments. This method allows us to provide the SCAPE Preservation Platform in the form of reusable images that can be published and launched on-demand using a private and/or public infrastructure like Amazon EC2¹. Virtual machine images have been proven to be particularly useful for packaging complex systems which require a tedious installation and configuration procedure, if being installed manually. The SCAPE Platform is a complex distributed system that requires considerable effort and experience in order to be installed and configured. Provided as pre-configured images, the Platform can be easily deployed on an arbitrary number of nodes, either manually or automatically based on cloud services (see section 5).

3.2 Employing a Cloud Hosting Model

The SCAPE Preservation Platform employs cloud technology on multiple levels includes hosting, computation, and storage. Infrastructure-as-a-Service (IaaS) provides a cloud hosting model that is well suited for an automated and scalable deployment of the platform environment. Figure 2 shows a simplified setup that utilizes a private cloud platform, as for example provided by the Eucalyptus [6] environment, to host an instance of the preservation platform. For simplicity, we assume that the cloud platform provides only two services, here called Cloud Controller (CC) and Storage Controller (SC). The cloud controller is capable of deploying virtual machine images on top of the cloud nodes (represented by the solid bordered boxes in the figure). The storage controller provides a service to store and retrieve the virtual machine images. Using this model, it is possible to bring up an instance of the preservation platform in a specific configuration. A platform instance comprising a (potentially large) number of platform nodes (PT(x)). Different nodes may have their own roles and behave differently within the platform instance. However, except nodes that provide external services, the platform internals are not visible to users/administrators.

¹<http://aws.amazon.com/ec2/>

3.3 Storing Data in the Cloud

When operating a computer cluster, frequently occurring node failures are an expected rather than an exceptional case. Consequently, the number of nodes within a cluster may dynamically grow and shrink over time. File systems like HDFS² can deal with such behavior by replicating and dynamically recovering data between the nodes. A virtual machine instance that might have been deployed in a cloud environment operates on a transient file system by default. This is to say that data that is stored using the virtual machine's file system will be erased once the instance is shut down. For its storage network, the SCAPE preservation platform, however, demands a persistent storage media, mainly for two reasons. (a) We expect that most deployments will be hosted in (research departments of) individual institutions rather than in large data centers. Here, it cannot be guaranteed that the IT-infrastructure can be kept-up-and-running over very long periods without any interruptions (e.g. caused by maintenance work). Using transient storage, it would however be virtually impossible to shut down the entire cluster without losing the data it holds. (b) A major design goal of the system is to support reconfiguration by deploying software environments, like specifically configured platform nodes, on demand. It is therefore required to separate the software environment's internal file system from the medium used to store content on the platform. The employment of a network attached storage system, however, would not satisfy the platform's scalability requirements, which demand to store data on a storage medium that is local to the processing unit of the node. As illustrated in figure 2, this can be solved by providing cloud nodes that provide a persistent storage layer upon which a software environment can be deployed dynamically (and operated locally to the data). While it is possible to establish such a configuration in a private cloud setting, this is usually not supported by commercial cloud offerings. Cloud storage is commonly provided as shared services that must be accessed via a network connection. The instantiation of environments on distinct physical computer nodes is in fact contradicting the public cloud model and can usually only be realized in a private setup.

4. APPLYING DATA-INTENSIVE TECHNOLOGIES

Preserving large volumes of loosely structured objects like data from scientific instruments, digitized objects, or multimedia content provides a resource demanding challenge. Both, scalable storage and processing capabilities are required to manage such data sets. In recent years, technologies like scalable file-systems, distributed databases, and frameworks for efficiently processing large quantities of data have emerged. We argue that the employment of scalable technologies can address and significantly enhance performance limitations of existing digital object management and preservation systems. These technologies were initially developed to capture and analyze vast amounts of data generated by Internet applications. Examples are data sets produced by social networks, search engines, or sensor networks. Such data sets have exceeded data volumes that can be organized using traditional database management tools. MapReduce [3] provide a prominent example of a distributed framework that

²<http://hadoop.apache.org/hdfs/>

is capable of processing huge amounts of data on top of a distributed file system. The MapReduce paradigm has also proven to be applicable to a range of domains. The SCAPE preservation platform is a software project that aims at employing scalable data management techniques for the purpose of digital preservation.

5. PRIVATE-CLOUD SETUP

AIT has started to deploy an initial version of the SCAPE Preservation Platform within a private cloud environment. The SCAPE infrastructure provides a Fully Automated Installation (FAI) server for configuring the cloud nodes. FAI is an automated installation framework that can be used to install Debian systems on a cluster. The service allows us to easily add new nodes to the system, which can be booted via a network card using PXE, a pre-boot execution environment most modern network cards support. The cloud infrastructure, presently consisting of 20 nodes, has been set up using the Eucalyptus cloud software stack. Eucalyptus is a private cloud-computing platform that provides REST and SOAP interfaces which are compliant with Amazon's EC2, S3, and EBS services. The infrastructure's front-end hosts the Eucalyptus Cloud Controller, the Cluster Controller, and the Walrus storage service. The worker nodes in the cloud run the XEN hypervisor and a Debian distribution that includes a Xen *Dom0* kernel.

The initial preservation platform is based on an Apache Hadoop³ cluster running MapReduce and HDFS, a set of preservation tools, and a number of MapReduce programs that have been developed to execute tools and/or specific workflows against data sets on the cluster. Using the cloud environment, a platform instance can be brought up dynamically by specifying a particular virtual machine image and the desired size of the cluster. The deployment of the platform instance supports the previously described requirements, namely dynamic deployment of environments and persistent storage. Each cloud node is configured with a physical data partition that can be hooked into the file system of a virtual machine instance. The platform nodes utilize this mechanism to establish a distributed file system that uses physical file system partitions underneath. Since data is already replicated by the Hadoop file system, it is not required to employ additional data redundancy mechanisms like RAID.

6. RELATED WORK

The employment of distributed and replicated storage and/or computation is a design decision that has been taken by a number of preservation systems. Prominent examples for systems that support geographically distributed and replicated data are LOCKSS [5] and iRods [7]. Preservation services like those developed in the context of the Planets project [8], provide a model that can be used to evaluate preservation tools in distributed environments. Many data management systems have been extended and/or configured to operate in cloud-based hosting environments. DuraCloud⁴ and Fedorazon⁵ are examples for repositories that leverage distributed cloud storage. The SCAPE platform

³<http://hadoop.apache.org/>

⁴<http://www.duracloud.org/>

⁵<http://www.ukoln.ac.uk/repositories/digirep/index/Fedorazon>

is intended to support existing digital object repositories and preservation environments. It leverages data-intensive computing techniques to achieve scalability regarding storage, throughput, and computation allowing users to perform preservation actions and data analysis tasks at scale. Cloud and virtualization technologies are employed to support dynamic reconfiguration of the specific software environment required to interpret and manipulate a particular data item closely to its storage location.

7. CONCLUSION

The architecture of the SCAPE Preservation Platform aims at a versatile design that is intended to be applicable to digital content from many domains and to different preservation and information management systems. This paper discusses general design decisions and provides an overview of possible hosting models. We conclude that a private cloud environment, as described in this paper, can provide a very powerful, secure, and versatile solution for hosting the preservation platform in an institutional environment.

Acknowledgments

Work presented in this paper is primarily supported by European Community's Seventh Framework Programme through the project SCAPE under grant agreements No 270137.

8. REFERENCES

- [1] ASSEG, F., RAZUM, M., AND HAHN, M. Apache Hadoop as a Storage Backend for Fedora Commons. In *7th International Conference on Open Repositories* (Edinburgh, UK, 2012).
- [2] BECKER, C., KULOVITS, H., GUTTENBRUNNER, M., STRODL, S., RAUBER, A., AND HOFMAN, H. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *Int. J. on Digital Libraries* 10, 4 (2009), 133–157.
- [3] DEAN, J., AND GHEMAWAT, S. MapReduce: Simplified Data Processing on Large Clusters. *Commun. ACM* 51 (January 2008), 107–113.
- [4] KING, R., SCHMIDT, R., BECKER, C., AND SCHLARB, S. SCAPE: Big Data meets Digital Preservation. *ERCIM News* 2012, 89 (April 2012).
- [5] MANIATIS, P., ROUSSOPOULOS, M., GIULI, T. J., ROSENTHAL, D. S. H., AND BAKER, M. The LOCKSS Peer-to-Peer Digital Preservation System. *ACM Trans. Comput. Syst.* 23, 1 (Feb. 2005), 2–50.
- [6] NURMI, D., WOLSKI, R., GRZEGORCZYK, C., OBERTELLI, G., SOMAN, S., YOUSEFF, L., AND ZAGORODNOV, D. The Eucalyptus Open-Source Cloud-Computing System. In *9th IEEE/ACM International Symposium on Cluster Computing and the Grid* (2009), CCGRID '09, IEEE, pp. 124–131.
- [7] RAJASEKAR, A., WAN, M., MOORE, R., AND SCHROEDER, W. A Prototype Rule-based Distributed Data Management System. In *in: HPDC workshop on "Next Generation Distributed Data Management, Paris, France.* (2006).
- [8] SCHMIDT, R., KING, R., JACKSON, A., WILSON, C., STEEG, F., AND MELMS, P. A Framework for Distributed Preservation Workflows. In *Proceedings of The Sixth International Conference on Preservation of Digital Objects (iPRES)* (San Francisco, USA, 2009).

Towards a Long-term Preservation Infrastructure for Earth Science Data

Arif Shaon, David Giaretta¹, Esther Conway, Brian Matthews, Shirley Crompton
Science and Technology Facilities Council Rutherford Appleton Laboratory Didcot, UK
{arif.shaon, david.giaretta, esther.conway, brian.matthews, shirley.crompton}@stfc.ac.uk

¹also Alliance for Permanent Access, director@alliancepermanentaccess.org

Jinsongdi Yu,
Jacobs University
Campus Ring 1, 28759 Bremen
Germany
j.yu@jacobs-university.de

Fulvio Marelli
European Space Research Institute/European Space Agency
Via Galileo Galilei, Casella Postale 64 Frascati (Rome), Italy
fulvio.marelli@esa.int

Ugo Di Giammatteo
Advanced Computer Systems
Via della Bufalotta 378, Rome, Italy
udig@acsyis.it

Yannis Marketakis,
Yannis Tzitzikas
Foundation for Research and Technology - Hellas (FORTH)
Institute of Computer Science
N. Plastira 100
Heraklion, Crete, Greece
{marketak, tzitzik}@ics.forth.gr

Raffaele Guarino
Capgemini
Rome, Italy
raffaele.guarino@capgemini.com

Holger Brocks
InConTec GmbH
Kirschenallee 7 96152 Burghaslach, Germany
holger.brocks@incontec.de

Felix Engel
FTK Association Research Institute for Telecommunications and cooperation Martin-Schmeißer-Weg 4, Dortmund, Germany
fengel@ftk.de

processing sources code, calibration tables, databases and ancillary datasets.

To maximise the value of ES data, its usage should not be limited to the domain of the scientists who originally produced it. ES data as a “research asset” should be made available to all experts of the scientific community both now and in the future. The ability to re-purpose existing ES data could cross-fertilise research in other scientific domains. For example, if epidemiologists can correctly interpret environmental data encoded in an unfamiliar format, the additional knowledge may assist them with understanding patterns of disease transmission.

Unfortunately getting access to all the necessary data and metadata is a serious problem; often the data are not available, accessible or simply cannot be used since relevant information explaining how to do so or the necessary tools, algorithms, or other pieces of the puzzle are missing. Moreover the ES data owners are dealing with the preservation and access of their own data and this is often carried out on a case by case basis without established cross-domain approaches, procedures and tools.

The SCIDIP-ES project¹ is developing services and toolkits which can help any organisation but the prime focus in this project is to show their use in ES organisations working with non-ES organisations concerned with data preservation to confirm the wide effectiveness in helping to improve, and reduce the cost of, the way in which they preserve their ES data holdings. In the following we describe how these services and tools are used to help to overcome some of the aforementioned problems faced by both the curators and the users of ES data, but it should be remembered that they are designed for much wider applicability.

In this paper, we discuss the key technical challenges and barriers of long-term ES data preservation that the SCIDIP-ES project is aiming to address. In addition, we highlight some examples

¹ The SCIDIP-ES project - <http://www.scidip-es.eu/>

gathered from the ES community during the first year of the project and present the SCIDIP-ES services and toolkits as solution to these community generated requirements.

2. BARRIERS AND CHALLENGES OF ES DATA PRESERVATION

The SCIDIP-ES project identified the following challenges based on the results of a series of surveys on various aspects of preserving ES data within the project, as well as related external materials, such as the PARSE.Insight case studies on the preservation of Earth Observation (EO) data [10]. Notably, some of the issues outlined here are also relevant beyond the ES and EO domains to the wider data preservation problem.

2.1 Ensuring Intelligibility and (Re-) Usability of Data

A frequently repeated mantra for digital preservation activities is “emulate or migrate”. However, while these activities may be sufficient for rendered objects, such as documents or images, they are not enough for other types of digital objects. In addition, there is a need to capture Representation Information (RepInfo) - a notion defined by the widely adopted ISO standard² Open Archival Information Systems (OAIS) Reference Model [1] to represent the information needed to access, understand, render and (re)use digital objects. The key aspects of RepInfo needed to ensure continued intelligibility and usability of data include Semantic Representation Information (i.e. intended meaning and surrounding context of data) and the identification of a Designated Community (consumer of the data).

Take for example some fairly simple tabular scientific data in an Excel spreadsheet. This can be easily migrated (or more accurately “transformed” in OAIS terms) to a comma-separated values (CSV) file. However if the semantics, such as the meaning of the columns and the units of the measurements is not recognised as important and preserved then the data will become meaningless and scientifically unusable. The problem is even more important for complex scientific data. Emulation to enable the continued use of the software used to handle the digital objects may be adequate for rendering these objects or re-performing previous operations. However, to combine the preserved data with newer scientific data will, in general, not be possible. For example, one may use an emulator to continue using the Excel software which has the semantics of what the columns mean encoded in its formulae, but one will not be able to combine this data with newer data, for example in NetCDF format³ which is a commonly used ES data format. Since emulators are a type of RepInfo, one can re-state the mantra as “collect RepInfo or Transform”.

This means that a key problem we need to address is – *how does a repository create or collect enough RepInfo?* It is difficult enough to deal with the complex dependencies of an ES data format like NetCDF; when one then looks at the multitude of ES and other scientific formats, each of which may have a plethora of associated semantic RepInfo (thus forming a tree or network of RepInfo dependencies), the problem explodes! In general, an archive may, depending on its data holdings, need various such networks - both individual and related. Hence, there is a need for

a service and tools to help spread the load in creating and managing RepInfo networks in a preservation archive or repository.

2.2 Designing a cost effective preservation solution

Long-term preservation archives and repositories must plan responses to changes and risks of changes in an appropriate and cost-effective way. As discussed above there are many different types of preservation action/strategy which are equally valid and need to be considered when a preservation solution is formulated for a data collection. Archives need to be aware of, characterise and describe the main types of preservation action available to an archivist. They also need to appreciate the effect each type of action has upon a RepInfo Network, the risks, available modes of stabilisation as well as cost and benefits. Hence, there is a need for tools to help evaluate and balance costs and risks in a RepInfo network. In addition, they need to consider how more than one type of strategy can be employed as alternates in order to create the optimal balance of risk and usability of a preservation solution.

2.3 Reacting to changes in preservation requirements

As mentioned above, long-term data archives need to be able handle changes in preservation requirements by re-strategising when needed. It is well understood that hardware and software become unavailable but also the semantics of specific terminology change and the knowledge base of the Designated Community, as chosen by a repository, changes. All these changes must be countered if we are to preserve our digitally encoded information. Yet *how can any single repository know of these changes?* Significant effort (e.g. the preservation watch service of the SCAPE project⁴) is being put into technology watches for document and image format changes. It is more difficult for a single repository to monitor all possible changes, such as in terminological changes across a multitude of scientific disciplines, and to understand the ramifications of such changes. From this perspective, there is a need for services to spread the knowledge, risk and implications of such changes.

2.4 Maintaining Authenticity

It is important to guarantee within an archive that digital data is managed and maintained through proper tools by applying suitable plans in order to ensure the “authenticity” of the data. In the OAIS model, authenticity of digital object is defined as *“the degree to which a person (or system) regards an object as what it is purported to be. Authenticity is judged on the basis of evidence.”* [1]

In general, any process and transformation could have side effects on digital data and corrupt the usability and integrity of the information being preserved. Therefore, authenticity requires more than just digital digests (e.g. checksum) – because these cannot by themselves guarantee that the data has not been altered, by accident or on purpose, by those in charge of the data and digests. Moreover the data may have been transformed from one form to another over time for a variety of reasons – the bit sequences and therefore the digests will change. More generally authenticity is not a yes/no issue – such as “does the digest match or not” – but rather a degree of authenticity judged on the basis of

² ISO 14721:2003 - http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683

³ NetCDF - <http://www.unidata.ucar.edu/software/netcdf/docs/>

⁴ The SCalable Preservation Environment (SCAPE) project - <http://www.scape-project.eu/>

technical and non-technical evidence. In effect, this involves capturing and evaluating that evidence as it is generated in many different ways over an extended time. Performing these tasks manually is likely to be laborious and even erroneous. This underlines the need for suitable tooling to facilitate capturing and evaluating the evidence needed to guarantee authenticity of data in a digital preservation archive.

2.5 Supporting Practical Business Models for Data Preservation

Preservation of data requires resources and long term commitments; therefore we need practical business models in order to build business cases for well identified “research assets” to justify their continued funding. At the same time the costs of preservation must also be reduced by avoiding unnecessary duplication of effort and wasting of resources, including energy. For instance, it may be financially more viable to turn an existing storage system into a preservation archive by integrating preservation services and tools into the existing system than to create a separate preservation archive.

However, no organisation can guarantee its ability to fund this storage and those responsible for the data will change over time. Long-term sustainability requires more than good intentions. It requires funding, and the recognition that the costs must be shared wherever possible. It also requires one to be realistic and recognise that no one repository can guarantee its existence forever; one must be prepared to hand over the digital holdings in a chain of preservation that is only as strong as its weakest link – and the hand-over from one link to the next must be easy and flawless. This hand-over is not just transfer of the bits but also the information which is normally held tacitly in the head of the data manager or embedded in the host data management system. We envisage that suitable and efficient services and tools can help prepare repositories for the hand-over process and moreover share the results and experience with the wider preservation community.

3. KEY USE CASES CONSIDERED IN SCIDIP-ES

The SCIDIP-ES project has defined the following three high level use cases to represent the main challenges of long-term preservation of ES data discussed above.

- **Preservation Archive Creation:** identifying what kind of information should be properly preserved for future use, by an identified Designated Community (DC) and the correct procedures needed to implement it. For existing archival systems, this would also need to address the efficient integration of preservation processes within the underlying system architecture.
- **Archived Data Access:** to add value to the preserved data, what kind of enhanced information could be provided to current and future consumers? In particular how can the repository enable a broader set of users to understand and use its data, e.g. to build a broader ES community, beyond the initial DC.
- **Archive Change/Evolution:** how to preserve data against changes in related technology (e.g. hardware, software) and in the designated community (data producer, data preserver, data consumer, the communities and organization involved in the information’s creation and initial use).

In this section, we describe the first high level use case – Preservation Archive Creation– with a specific focus on the ESA

ENVISAT MERIS dataset⁵. This is because the Archive Creation/Enhancement use case is the milestone on which the following use cases are built.

3.1 Preservation Archive Creation

We have defined the following logical model as a guideline to structure the archive definition phase workflow (Figure 1).



Figure 1. Phases of Preservation Archive Creation

3.1.1 Define the Preservation Objective

A preservation objective defines the minimum level and type of reuse which an archive wishes to maintain for its user community. Typical this would cover areas such data processing, visualization, analysis and interpretation of data. For example, MERIS has provided ten years of detailed observations of land, atmosphere and oceans. The Objective is to preserve ESA’s MERIS data package to maintain its time series, accessible and usable by different scientific user communities for 50 years. The minimum guaranteed level of preservation is the storage/archiving of the **ESA MERIS N1 File Level 0 (L0)** and **Level 1 (L1)**. The L0 data is the lowest level product and derived from MERIS. It is the satellite raw data which has been simply reformatted and time ordered in a computer readable format. L1 is derived from L0 data and both use the N1⁶ file format. L1 data, among other processes, is geo-located, calibrated and separated from auxiliary data. We focus in this section on preservation of L0 and will discuss the preservation of L1 data in subsequent sections.

3.1.2 Definition of the Designated Communities

The definition of the DC should specify the skills, resources and knowledge base a community has access to. DC description must have sufficient detail to permit meaningful decisions to be made regarding information requirements for effective re-use of the data. In the MERIS case, the DCs (both archive and user community) include:

- ESA staff – with full specific knowledge of ENVISAT datasets management.
- Principal Investigator (PI) - working on Earth topics such as Agriculture, Atmosphere, land, Natural disaster, Ocean, etc. They know the ENVISAT data scientific value but don’t have the skills to manage it.
- University Students - they are learning ENVISAT data and need to fully understand and use it.

⁵ ENVISAT Meris Instrument description and access to data can be found at <https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/envisat/instruments/meris>

⁶ The N1 File Structure - <http://www.noc.soton.ac.uk/bilko/envisat/tutorial/html/t0110.html#sh2>

3.1.3 Preserved Dataset Content Definition

Once the objectives and communities have been identified and described, an archive should be in a position to determine the full set of information required to achieve an objective for this community. To allow processing, visualization, analysis and interpretation of ESA MERIS data and the correct utilization by anyone with basic knowledge of the EO domain, the Archive must contain comprehensive information about:

- Science Data Records: raw data, L0 and L1 data, browse images, ancillary data, auxiliary data , Calibration and Validation data
- Processing software and databases: L0 consolidation software, instrument processing software, quality control software, data visualization tools
- Mission Documentation

3.1.4 Create Inventory

The next stage is to appraise each of the information objects in terms of physical state, location and ownership. The resulting inventory should include details of each of the pieces of Information, its Location, Physical State and associated Intellectual Property Rights (IPR). For example, the MERIS inventory would contain MERIS processing software and databases including:

- L0 consolidation software (mission dependent) described in the mission products description document. This document is available and is the IP of ESA.
- The Basic ENVISAT Toolbox developed to facilitate the utilization, viewing and processing of ENVISAT MERIS data along with the associated GNU public license
- The Java Virtual Machine required to run the ENVISAT Toolbox; Oracle owns several aspects of JAVA related IP.

3.1.5 Perform Risk Assessment

There may be a number of key risks associated with the MERIS data as described in the following categories and examples:

Technical Risk: software for processing the MERIS data (e.g. BEAM software) run with specific libraries (e.g. JVM1.5). Thus, it is also necessary to preserve such information so that the whole chain of soft/hardware dependencies could be evaluated.

Organizational Risks: ESA may decide to store copies of the MERIS data in different geographical locations to safeguard the archive from external hazards like floods and other natural disasters or technological hazards, etc.

IPR related Risks: As a research organization, ESA encourages, protects and licenses innovations or original works resulting from its activities. The MERIS data is protected according to the ESA IPR guidelines⁷. The need is to ensure that IPR or licences related to data, software (e.g. BEAM) and libraries (e.g. Java 1.5) are assessed for potential breaches.

Resourcing Risks: The preservation plan exists on the basis that funding and skills to support the data archive will be available for a defined time period. Should any of this change, the plan will need to be adapted.

3.1.6 Preservation Planning and Risk Monitoring

Preservation planning is the process which designs the long term research asset to be preserved within an Archival Information Package (AIP). AIP conceptually contains all the information required to ensure the long term usability of digitally encoded information. The cost, benefits and risk burden acceptable to an archive will determine the optimal preservation action to adopt. Preservation actions for construction and maintenance of the AIP take one of the following forms: *Risk Acceptance and Monitoring (referencing)*, *Software Preservation or Description and Transformation*.

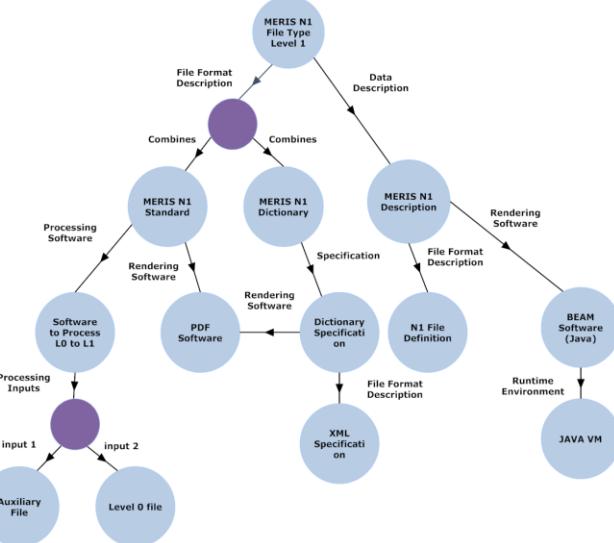


Figure 2. Network of RepInfo for ENVISAT MERIS data

Using the notion of Preservation Network Model described in Section 4.1.4, we designed a network of RepInfo (Figure 2) for the MERIS L1 example. This defines the whole chain of dependencies required to preserve its data and the associated knowledge to interpret it. As no preservation solution is permanent or necessarily stands the test of time, AIPs must be monitored for stability and suitability. To achieve this, the accepted risks/dependencies within the preservation network as well as the preservation objective and DC description must be recorded and monitored.

4. SCIDIP-ES PRESERVATION INFRASTRUCTURE

To address the long-term preservation challenges of the ES data (Section 2), in SCIDIP-ES, we aim to put in place an e-infrastructure consisting of various services and toolkits to facilitate long-term data preservation and usability. In essence, we combine a top-down, data centric view, using a proven design for generic infrastructure services to enable persistent storage, access and management, with a bottom-up, user-centric view, based on requirements from the ES community. The former comes from leading research projects in digital preservation, in particular CASPAR. The latter is from the developing European Framework on Long Term Data Preservation (LDTP, coordinated by ESA) for Earth Observation data.

⁷ ESA Intellectual Property Rights - http://www.esa.int/esaMI/Intellectual_Property_Rights/index.html

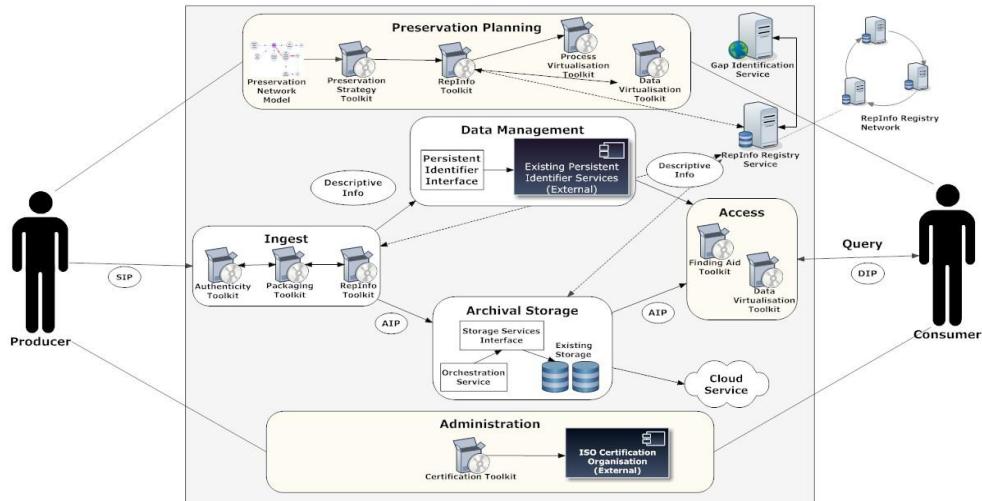


Figure 3. Overview of the services and toolkits within the SCIDIP-ES preservation infrastructure

4.1 SCIDIP-ES Preservation Services and Toolkits

To ensure consistency and interoperability, we use the OAIS Reference Model to underpin the definitions of the services and toolkits of the SCIDIP-ES preservation infrastructure. While the infrastructure is intended to cover the full preservation workflow defined in the OAIS model, the specific areas focused on are those connected with the construction of AIPs. As discussed earlier in the paper (Section 2.5), no organisation can be expected to look after a piece of data forever but rather that it can hand on its holdings to the next in the chain of preservation. Such a process can be hindered by lack of clear understanding of tacit dependencies and knowledge, and insufficient time available during the hand-over to capture these. Creation of an AIP ensures that these are made explicit well before they are needed, and so any future hand-over can be smooth and complete.

The SCIDIP-ES preservation infrastructure consists of the following services and toolkits (as shown in Figure 3) that have been defined to support both the data and user centric views that we have adopted. It enables the ES repositories to effectively address the challenges of preserving ES data mentioned in Section 2. A logical overview of these services and how they support the stages of the OAIS reference model is given in section 3.

4.1.1 RepInfo Registry Service

The RepInfo Registry Service is essentially a web-service based repository that is used to store, query, retrieve and manage the RepInfo needed to enable access, understanding and (re-)use of a digital object over the long-term. The RepInfo provided by the Registry Service can cover the structure of the digital object (format, headers, footers, instrument measures, annotations, fixed parts, variable parts, etc.), the semantics of that digital object (semantics, auxiliary information, usage information), and other information (e.g. rendering information which describes what additional software can be used to display/process/edit the digital object).

The Registry will also enable users to navigate a RepInfo network to explore the knowledge represented (e.g. a satellite image is linked to the image sensor description, which is linked to the satellite mission description, etc.).

4.1.2 The RepInfo Toolkit

The RepInfo toolkit provides a user-friendly GUI to the Registry to enable various components to interact efficiently with it. For example, the ingest and planning of the preservation life cycle in an archive. It also provides users with a set of tools to create the RepInfo required for specific digital objects. Some sub-components of this toolkit are aimed at describing the data in more “virtualised” terms which can help integrate data into other software.

4.1.3 Gap Identification Service

As underlined in the OAIS model, there is a need for services that help archivists in checking whether the archived digital artefacts remain understandable, and to identify hazards and the consequences of probable losses or obsolescence risks. In SCIDIP-ES, we have defined the Gap Identification Service (GIS) to facilitate such assessments of intelligibility of digital objects by identifying “gaps” in the corresponding RepInfo Network in the RepInfo Registry. In essence, this service is inspired by a model that consists of the notions of module, dependency and profile as discussed in [3]. If applied to digital objects, a module can be a software/hardware component or even a part of the knowledge base expressed either formally or informally, explicitly or tacitly, that we want to preserve. The dependency is captured in the logical links in meaning between modules. In addition, a module may require the availability of other modules in order to function, be understood or managed (e.g. a network of RepInfo). A profile is the set of modules that are assumed to be known (available or intelligible) by a user (or community of users), so this is an explicit representation of the concept of Designated Community Knowledge Base (KB). Utilising this model, the GIS is able to check whether a digital object (module) is *intelligible by a community*, and to compute the *intelligibility gap* of a digital object.

In an archive, the GIS can be used in the preservation planning process to evaluate the current knowledge base of the designated community as well as future review(s) of the plans by analysing changes in the related knowledge base. It can also be used for deriving DC-aware AIPs or DIP (dissemination information packages).

4.1.4 Preservation Strategy Toolkit

There are a number of basic strategies for preserving digitally encoded information. Besides describing the data using RepInfo,

one could transform the data into a different format or emulate the essential software to access the preserved information. The Preservation Strategy Toolkit helps repositories decide which technique to use, balancing costs against efficacy for given specific preservation objectives.

The toolkit uses the Preservation Network Model (PNM - see Figure 2 for an example), which was developed within the CASPAR project in order to represent the output of a preservation analysis conducted for a digital object to be preserved in a preservation archive or repository [4]. The preservation analysis of a digital object enables identification and assessment of the risks associated with its dependencies on other entities. The output of this type of analysis underpins the formulation of a suitable preservation strategy to be adopted by an archive; taking into account the preservation aims, related risk tolerance level, preservation policies and other requirements. The PNM can be used to articulate the result of preservation analysis as a network of related objects along with the preservation decisions associated with the relationships between the objects.

In an OAIS-compliant archive, the use of this toolkit would be a part of the **preservation planning** process/stage (see section 3.1.6), where the toolkit would also need to interact with the RepInfo toolkit to query and retrieve existing RepInfo records and/or create new ones as determined by the planning process.

4.1.5 *Authenticity Toolkit*

The Authenticity Toolkit is used to capture appropriate evidence of the authenticity of the digital object including that obtained from the Submission Information Package (SIP) during Ingest. As defined in the OAIS model, this authenticity evidence forms the Preservation Description Information (PDI) about the digital object and consists of various types of information including Reference, Context, Provenance, Fixity and Access Rights. The main underlying idea is to help to ensure that appropriate provenance is captured, for example if the data is transformed to a different format. Provenance is used to assess the Authenticity of a particular digital object.

4.1.6 *Packaging Toolkit*

The Packaging Toolkit is used (mainly during Ingest) to construct AIPs that will be stored using the Storage Service (see Section 4.1.7). The information collected in an AIP is aggregated either physically, or more likely, logically. In the latter case, the toolkit identifies within the AIP the location of the components so that they can be instantiated as a physical object for dissemination when requested by user. Additionally, the packaging toolkit needs to interact with the RepInfo toolkit to identify and obtain the RepInfo to accompany the digital object in an AIP.

4.1.7 *Storage Services*

This service provides an interface to the physical storage of digital objects. Using this interface ensures that all the information needed for the long term preservation of the data is identified (in an AIP) and can be moved from a repository to another when, for example the funding for the former ends. The interface can be implemented on top of existing storage systems so there should be no need to make major changes in existing repositories – it just adds the AIP capabilities. New storage systems could also be adopted, though this would not be without costs. For example in the last years, storage services have been progressively moving to web-based platforms in which the user sees a virtual archive that seamlessly takes care of all storage

functions (data distribution, redundancy, refresh, etc.). Cloud storage is the technological basis for this service, which hides the physical storage complexity.

Therefore, in the SCIDIP-ES project, the aim is not to develop a new storage service for the ES data but to provide the data holders with “preservation-aware” storage service infrastructure based on existing storage technologies including cloud-based services.

4.1.8 *Persistent Identifier Service*

The ability to unambiguously and persistently locate and access digital objects is an important requirement of successful long-term digital preservation. In a digital preservation archive, the use of persistent identifiers (PIs) is ubiquitous including identifying AIPs in the storage as well as the RepInfo records in the RepInfo Registry. Assigning PIs to objects is usually the task of the **Data Management** component ([1]) of the archive.

In SCIDIP-ES, we aim to develop a simple persistent identifier service that interfaces to multiple existing Persistent Identification (PI) systems (e.g. DOI⁸) to obtain a unique identification code for the digital objects that are created within the system. It allows the interoperation of persistent identifiers used in different repositories and spreads the risk associated with a single PI system.

4.1.9 *Orchestration Service*

The Orchestration Service provides a brokerage service between existing data holders and their successors. Additionally, it also serves more generally as a knowledge broker. In particular it can exchange intelligence about events which might impact the long-term usability and/or access of data, e.g. changing technologies (support for new media and data formats), changing terminologies/knowledge of the DC and even changing ownership of data/ archive. Each of these kinds of changes may bear certain preservation risk concerning the data holdings in question. The Orchestration Service is intended to act as a collector of information about these kinds of events and broker the corrective actions necessary.

4.1.10 *Finding Aid Toolkit*

To support users' need to access and use data from many sources across many domains the infrastructure will provide a Finding Aid Toolkit to supplement the many existing domain search facilities. The development of this toolkit will aim to address, by utilising and harmonising related metadata and semantics (ontologies), the discovery of ES data that are not easily discoverable and accessible as they are heterogeneous in nature, (e.g. data coming from different sensors on different platforms such as satellites, aircraft, boats, balloons, buoys or masts, or located on the land), they are spread all over the world and originate from different applications.

This toolkit is not strictly related to data preservation process but it is a fundamental instrument to allow digital objects to be discovered by users and can play a fundamental role when it comes to data interoperability between different user communities.

4.1.11 *Data Virtualisation Toolkit*

The Data Virtualisation Toolkit allows the curators to inspect and describe the contents and structure of a digital object in a format independent manner creating the appropriate RepInfo. For example, in principle, using the toolkit, the contents of a

⁸ Digital Object Identifier - <http://www.doi.org/>

NetCDF-based file could be viewed in a tabular format without needing a dedicated NetCDF viewer. In addition, the toolkit could also be used to help create (using the RepInfo toolkit) further RepInfo about any sub-components of the object as part of preservation planning and analysis. We envisage that this could also facilitate data access - i.e. consumers could use this type of RepInfo to bring together and analyse data from multiple sources without having to use multiple dedicated software systems. If full analysis capabilities are not available in this way the consumer could inspect the actual content of a digital object before making the effort to obtain all the RepInfo needed to use it.

4.1.12 Process Virtualisation Toolkit

Process Virtualisation Toolkit is of fundamental importance in cases where digital objects need to be re-processed in the future to generate added value products. Thus, all information and/or ability to perform the digital object processing need to be preserved as well. The process virtualisation describes, in general terms, the various processes associated with the data by enabling an archivist or repository manager to identify the missing pieces of a given processing chain and apply corrective actions. For example, it may be necessary to re-compile the source code in order to run it in a different infrastructure ("create L-1C product from L-1B and port to new processing environment") as well as instantiating virtual host on-demand for processing.

4.1.13 Certification Toolkit

The Certification Toolkit will be a relatively simple tool for collecting evidence based on the ISO 16363⁹ draft standard to submit for the ISO certification process. In addition, this tool will also be useful for self-audit and preparation for full, external audits.

4.2 Initial Prototypes and Validation

In the initial phases of SCIDIP-ES, we have developed basic prototypes of six of the services and toolkits - RepInfo Registry, Gap Identification Service, RepInfo Toolkit, Packaging Toolkit, Orchestration Manager and Data Virtualisation Toolkit. The development of these initial prototypes have been based on their original implementations by the CASPAR project, which also produced an extensive collection of evidence of their effectiveness in terms of preservation in several science disciplines (STFC and ESA repositories), cultural heritage (UNESCO world heritage¹⁰) and contemporary performing arts (INA, IRCAM, ULeeds and CIANT).

In SCIDIP-ES, these prototypes have undergone further evaluation by the projects partners, in particular the ones with ES data holdings with a view to understanding how the prototypes would be used in their archives. As a key outcome, this evaluation identifies the need for the services and toolkits to be more robust, scalable and simplified, where possible. These prototypes are publicly accessible for review¹¹.

4.3 Key Implementation Challenges and Future Work

As mentioned above, the majority of the services and toolkits in SCIDIP-ES were originally designed and implemented as proof-

of-concept prototypes by the CASPAR project. In SCIDIP-ES, we aim to turn the CASPAR prototypes into production quality services, that is operational, scalable and robust as well as simplified (where possible) software products. In the process we will re-design the specifications based on the user cases and requirements defined in the project.

To address the scalability requirement of the RepInfo registry service, we aim to develop a network of RepInfo Registry services, section 4.1.1 to enable load distribution of requests for RepInfo between multiple registries acting as "Nodes" in the network. In order to avoid a single point of failure, all the registries will be essentially identical, apart from their holdings of RepInfo. There will be at least one registry, the Guarantor Node, in the network which we guarantee will be running even if all the others close down. The name(s) of the Guarantor node(s) will be propagated (e.g. via configuration in registry.representation.info property) so that new registries can register themselves with it.

The Gap Identification Service needs to improve the speed of query processing and providing support for transitive queries (a rule-based approach is investigated in [11]), while the Orchestration Service requires improved and more efficient support for the notification of preservation related events.

For the toolkits, such as Authenticity, Provenance and Integrity Toolkit and the Preservation Strategy Toolkit, we will aim to incorporate scalability in the underlying information models. Our analysis indicates that scalability of this toolkit could be achieved by creating a PNM record per group or collection of digital objects rather than per digital object in an archive. We are also exploring the feasibility of profiling the PREMIS metadata model [5] in the form of an OWL ontology to enable automation of creation and management of PNM records.

In addition, the development of the Persistent Identifier Interface Service will collaborate with the APARSEN project¹² that is developing an interoperable framework for persistent identifier services. We will also leverage the work done by the SHAMAN project, particularly the SHAMAN Preservation Context Model [6], for addressing the scalability and other related issues associate with the Process Virtualization and Emulation, and Packaging Toolkit. Similarly, we plan to build on the CASPAR Preservation Data Store [7] interface and the Kindura project's [8] approach to integrating traditional data system with Cloud-based technology in order to develop "preservation-aware" interfaces to any suitable existing storage systems. In effect, this would serve as the Storage Services needed for the SCIDIP-ES services and toolkits.

Besides the scalability and robustness issues, we have a number of other challenges. The infrastructure and toolkits must be usable in a number of existing systems. For instance, we plan to build the data virtualisation toolkit for ES data as a uniform front-end on a variety of existing data specific tools, including those creating standardised data descriptions using specifications such as EAST¹³ and DFDL¹⁴. To verify this we will show effective integration in several different repositories, supporting their decision making and respecting their existing hardware, software, and governance and control systems. We must be able

⁹ ISO 16363 -
http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510

¹⁰ UNESCO world heritage - <http://whc.unesco.org/>

¹¹ <http://jenkins.scidip-es.eu/joomla/>

¹² The APARSEN Project -
<http://www.alliancepermanentaccess.org/index.php/aparsen/>

¹³ The Data Description Language EAST - A Tutorial -
<http://public.csds.org/publications/archive/645x0g1.pdf>

¹⁴ Data Format Description Language - <http://www.ogf.org/dfdl/>

to create RepInfo to enable broader use of the data. This will be verified as we enable ES users in different disciplines to find and use each other's unfamiliar data.

5. CONCLUSIONS

In this paper we have discussed the motivations and approach in the design of a preservation infrastructure, initially targeted at ES, but which has wider application across scientific disciplines. There are two aspects to this which we wish to highlight.

Firstly, that there needs to be a thorough preservation analysis to establish the context in which the preservation initiative takes place. For scientific data such as Earth Sciences this is an analysis beyond the digital objects themselves to consider both the dependencies on other entities that provide contextual information, which themselves have dependencies to form a network, and the requirements and assumptions of the designated community. From this analysis, an assessment of the costs and benefits of the preservation can be undertaken, taking into account the risks involved in preservation. In SCIDIP-ES we are demonstrating the value of this approach in practise in the implementation of the use cases, via the use of Preservation Network Models.

An important outstanding aspect of this approach is the establishment of the value of preservation for the archives involved. This is necessarily difficult to assess; it is particularly difficult to give the value of the use of data to support future scientific advances and subsequent impact on society. However, a framework for estimating the value proposition is nevertheless required to justify the additional effort of preservation, which does tend to be front-loaded. For archives such as those involved in SCIDIP-ES, which are from publically supported science, the framework should extend beyond the purely commercial to cover research and ultimately societal benefits. An ongoing work item within SCIDIP-ES is exploring such a framework.

Secondly, in order to support effective preservation, an infrastructure with a number of services needs to be provided to support the stages of the OAIS functional model. In this paper, we have outlined the services identified within SCIDIP-ES and discussed how they might interact to support a preservation scenario. To be sustainable, these services must be of production quality. We have discussed their scalability, in both size and heterogeneity. But to be production quality these services also need to be robust in the presence of failure, secure to maintain the integrity of the data, and maintain accessibility as the environment changes. Thus to be sustainable, these tools themselves need to adhere to a strong preservation discipline.

SCIDIP-ES has completed its initial analysis and design, as reported in this paper. In the next phases of the project, this sustainable infrastructure will be realised, deployed, and evaluated on the ES use cases.

Finally, it should be noted that the SCIDIP-ES preservation services and toolkits are designed for much wider application than the Earth Science use cases considered in this paper. For instance, we envisage that the SCIDIP-ES infrastructure, when developed, will have the potential to aid long-term preservation of data exposed through the large-scale Spatial Data Infrastructures (SDIs) in Europe, such as the INSPIRE SDI¹⁵, which currently do not address preservation [9].

6. ACKNOWLEDGMENTS

The work presented in this paper is funded by the Seventh Framework Programme (FP7) of the European Commission (EC) under the Grant Agreement 283401.

7. REFERENCES

- [1] CCSDS. 2002. Reference Model for an Open Archival Information System (OAIS). Recommendation for Space Data Systems Standard, *Consultative Committee for Space Data Systems (CCSDS) Blue Book*, 2002, or later URL: <http://public.ccsds.org/publications/archive/650x0b1.pdf>
- [2] Y. Marketakis and Yannis Tzitzikas 2009: Dependency management for digital preservation using semantic web technologies. Int. J. on Digital Libraries 10(4): 159-177 (2009)
- [3] Conway, E., Dunckley, MJ., Giaretta, D. and Mcilwrath, B. 2009. Preservation Network Models: Creating Stable Networks of Information to Ensure the Long Term use of Scientific Data, *Proc. Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data*, del Castillo, Madrid, Spain, 01-03 Dec 2009. URL: http://epubs.cclrc.ac.uk/bitstream/4314/PV09_Conway_PN_M.pdf
- [4] PREMIS. 2008. PREMIS Data Dictionary for Preservation Metadata, version 2.0, *PREMIS Editorial Committee*. URL: <http://www.loc.gov/standards/premis/v2/premis-2-0.pdf>
- [5] Brocks, H., Kranstedt, A., Jäschke, G., and Hemmje, M. 2010. Modeling Context for Digital Preservation. In *E. Szczerbicki & N. T. Nguyen (Eds.), Smart Information and Knowledge Management* (Vol. 260, pp. 197–226). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-04584-4_9
- [6] CASPAR Consortium. 2008. D2201: Preservation Data Store Interface, *CASPAR Consortium*, 2008. URL: http://www.casparpreserves.eu/Members/cclrc/Deliverables/preservation-datastore-interface/at_download/file.pdf
- [7] Waddington, S., Hedges, M., Knight, G., Zhang, J., Jensen, J. and Downing, R. Kindura. 2012. Hybrid Cloud Repository, *Presentation*, 2012. URL: http://www.jisc.ac.uk/media/documents/events/2012/03/waddington_kindura.pdf
- [8] Shaon, A., Naumann K., Kirstein M., Rönsdorf C., Mason P., Bos M., Gerber U., Woolf A. and G Samuelsson G. 2011. Long-term sustainability of spatial data infrastructures: a metadata framework and principles of geo-archiving, *Proc. 8th International Conference on Preservation of Digital Objects*, Singapore, 01-04 Nov 2011. URL: http://epubs.stfc.ac.uk/bitstream/7195/GeoPres_IPRES_CR.pdf
- [9] PARSE.Insight. 2010. Case Study Report, *PARSE.Insight Public Report*, 2010. URL: http://www.parse-insight.eu/downloads/PARSE-Insight_D3-3_CaseStudiesReport.pdf
- [10] Y. Tzitzikas, Y. Marketakis and Y. Kargakis, 2012, Conversion and Emulation-aware Dependency Reasoning for Curation Services, iPres 2012

¹⁵ INSPIRE - <http://inspire.jrc.ec.europa.eu/>

Migration at Scale: A Case Study

Sheila M. Morrissey

ITHAKA

100 Campus Drive, Suite 100
Princeton NJ 08540 USA
1-609-986-2221

sheila.morrissey@ithaka.org

Matthew Stoeffler

ITHAKA

301 East Liberty, Suite 250
Ann Arbor MI 48104 USA
1-734-887-7079

matthew.stoeffler@ithaka.org

Vinay Cheruku

ITHAKA

100 Campus Drive, Suite 100
Princeton NJ 08540 USA
1-609-986-2232

vinay.cheruku@ithaka.org

William J. Howard

ITHAKA

100 Campus Drive, Suite 100
Princeton NJ 08540 USA
1-609-986-2217

william.howard@ithaka.org

John Meyer

ITHAKA

100 Campus Drive, Suite 100
Princeton NJ 08540 USA
1-609-986-2220

john.meyer@ithaka.org

Suresh Kadirvel

ITHAKA

100 Campus Drive, Suite 100
Princeton NJ 08540 USA
1-609-986-2273

suresh.kadirvel@ithaka.org

ABSTRACT

Increasing experience in developing and maintaining large repositories of digital objects suggests that changes in the large-scale infrastructure of archives, their capabilities, and their communities of use, will themselves necessitate the ability to manage, manipulate, move, and migrate content at very large scales.

Migration at scale of digital assets, whether those assets are deposited with the archive, or are created as preservation system artifacts by the archive, and whether migration is employed as a strategy for managing the risk of format obsolescence, for repository management, or for other reasons, is a challenge facing many large-scale digital archives and repositories.

This paper explores the experience of Portico (www.portico.org), a not-for-profit digital preservation service providing a permanent archive of electronic journals, books, and other scholarly content, as it undertook a migration of the XML files that document the descriptive, technical, events, and structural metadata for approximately 15 million e-journal articles in its archive. It describes the purpose, planning, technical challenges, and quality assurance demands associated with digital object migration at very large scales.

Categories and Subject Descriptors

H.3.7 [Digital Libraries]: Language Constructs and Features – Collection, Standards, Dissemination, Systems issues.

General Terms

Management, Measurement, Documentation, Economics, Reliability, Standardization, Verification.

Keywords

Digital preservation, archives management, format migration, transformation, at scale, normalization.

1. BACKGROUND

1.1 Format migration

Increasing experience in developing and maintaining large repositories of digital objects suggests that changes in the large-scale infrastructure of archives, their capabilities, and their communities of use, will themselves necessitate the ability to manage, manipulate, move, and migrate content at very large scales.

Migration at scale of digital assets (whether those assets are deposited with, or created as preservation system artifacts by the archive) is therefore a challenge facing many large-scale digital archives and repositories. This is true whether migration (or, alternatively, “transformation”, or “normalization”) occurs at the point of ingest into the archive, at the point of delivery of a digital artifact from the archive, or as part of ongoing archive management.

There are many motivations for performing a format migration. It might be undertaken as part of a repository’s preservation strategy: to ensure access to a digital object in an obsolete or obsolescing format, or in conformance with a repository’s policy to support a fixed list of formats consider to be at a lesser risk of obsolescence [5]. It might be undertaken to replace or complement an archival master object with an instance in a more compact format, either to save on storage costs, or to reduce bandwidth and latency on a rendition version of the object [13]. It might be undertaken to create a “normalized” view of archive content, as an aid to search, discovery and management [1], or to establish whether later migration (whether for delivery or other reasons) is likely to encounter difficulties[2]. And it might be motivated by new developments, both in technology and in the requirements and expectations of (possibly new) communities of use, that result in new, and originally unanticipated, uses of content in repositories. Such, for example, would be the extraction of “text content” from non-text format instances (for example, constructing text content from instances of page image formats such as PDF and TIFF) across all instances of those formats in a repository, to facilitate large-scale content-mining of digital corpora.

This paper explores the experience of Portico as it undertook a migration of the XML files that document the descriptive,

technical, events, and structural metadata for approximately 15 million e-journal articles in its archive. It describes the migration purpose, planning, technical challenges, and quality assurance demands associated with digital object migration at very large scales.

1.2 Portico Preservation Workflow and Metadata

Portico is a digital preservation service for electronic journals, books, and other content. Portico is a service of ITHAKA, a not-for-profit organization dedicated to helping the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways. As of May 2012, Portico is preserving more than 19.4 million journal articles, e-books, and other items from digitized historical collections (for example digitized newspapers of the 18th century).

Content comes to Portico in approximately 300 different XML and SGML vocabularies. These XML and SGML documents are accompanied by page image (PDF, TIF, and JPG) and other supporting files such as still and moving images, spreadsheets, audio files, and others. Typically content providers do not have any sort of manifest or other explicit description of how files are related (which ones make up an article, an issue of a journal, a chapter of a book). This content is batched and fed into a Java workflow, called the “Content Preparation” (ConPrep) system, for assembly into what the Open Archival Information System (OAIS) Reference Model terms “Submission Information Packages” (SIPs) [3].

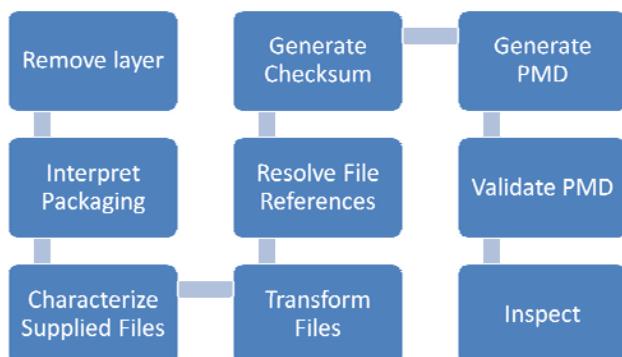


Figure 1 Portico ConPrep High-Level Workflow

The ConPrep workflow maps the publisher-provided miscellany of files into bundles that comprise a single article or book or other content item, which Portico terms an “archival unit” (AU). It identifies the format of each of the AU’s component files, and, where a format specification and validation tool is available, validates each file against its format specification. Publisher-provided XML and SGML journal article files are normalized to the Portico profile of the National Library of Medicine’s Journal Archiving and Interchange Tag Set; e-book files are normalized to a profile of the NLM’s NCBI Book Tag Set. So ConPrep, which has processed and packaged Portico’s archival units into SIPs, is itself an instance of migration at scale, of both the (implicit

package format and of files within the package, at the point of receipt of content.

Some of the steps in this workflow are automated quality-assurance checks of the XML content – both the content provided by the publishers, and artifacts produced by Portico in the workflow itself. This QA includes validation against XML and SGML document type definitions (DTDs) and schemas. It also includes the assertion, via Schematron (a rule-based validation language for making assertions about the presence or absence information in XML files [7]) of other constraints on content values. Additionally, the workflow includes visual inspection of sample content.

ConPrep generates preservation metadata for each AU. Modeled on PREMIS [10] and METS [4], the generated information includes descriptive, or bibliographic, metadata; structural metadata specifying the relationships among the components of the archival unit, technical metadata about files and their formats; provenance and event metadata, detailing the tool chain, including hardware and software information used in processing the content, and rights metadata stipulating Portico’s legal right to preserve these digital objects. These metadata are instantiated as XML, and are stored with the preserved digital object. Just like the publisher-provided XML files, the preservation metadata is schema-validated, and then further validated via Schematron.

2. THE MIGRATION

2.1 Motivation

2.1.1 Archive Life Cycle: Continual Review and Revision

As with its preservation policies, practices, and procedures, Portico’s preservation infrastructure – including its hardware, software, and key data and metadata structures – has been subject since inception to a continual process of review and revision. This review and revision is intended to incorporate both lessons learned from our own experience with content that has steadily expanded both in volume and in type, and with the continually developing understanding of best preservation practice in the larger preservation community.

The first major refinement of the original Portico platform was undertaken to scale up the capacity of the ConPrep system from 75,000 e-journal articles (and approximately 750,000 files) per month (900,000 articles/9,000,000 files per year) to 10 million articles and 100 million files per year – an order of magnitude increase. The system was in fact increased to a capacity of 24 million articles and 240 million files per year, operating at 50-75% of peak capacity. [11]

2.1.2 New Requirements, New Knowledge: New Content Model

As the Portico archive was extended to handle new content types beyond electronic journal content, its content model and the Portico metadata (PMETS) schema (which had key conceptual dependencies on that content model), were subjected to review and revision. The PMETS schema, whose design was based on METS 1.4, and informed by early work on the then-uncompleted PREMIS data dictionary, had undergone 6 minor, backwardly

compatible revisions (typically to accommodate changes to subsidiary schemas which specified descriptive and events metadata) since it was designed and implemented in 2002-2003.

By late 2008, the review process indicated the data model underlying the PMETS schema would be stressed by new requirements for the Portico archive. These included

- new content types (such as books and digitized collections), with richer and more complex relationships among the components comprising a single digital object
- new preservation activities, such as versioning, the creation of access artifacts, and the export of metadata in standard formats
- extended use cases in the ConPrep system, including the ability to assign preservation level by business policy rather than only by file format validity; to de-duplicate content in the archive; to process externally updated content (new versions of all or part of a content unit) as well as internally updated content (such as new technical metadata generated by newly available tools); to capture “use” information (for example, that information that one image file is a “thumbnail” of another image file); to record and manage migration and re-migration of content

The main components of the Portico content model (both the old and new versions) are:

- Content Type (CT) – This allows Portico to group content belonging to specific preservation services together, and allows us to group “like” objects together.
- Content Set (CS) – This allows Portico to group together archival units that belong together. For example, all archival units for a single journal of a particular publisher will be placed together within a single content set.
- Archival Unit (AU) – The main digital object or abstract intellectual object that is being archived. For example an E-Journal Article.
- Content Unit (CU) – A complete version of the content for an AU. In most cases, an AU will only contain a single CU.
- Functional Unit (FU) – A container for grouping together components that serve the same function within a content unit. For example, the high-resolution, web ready and thumbnail versions of an image for a single equation or chemical formula would be grouped together in a single FU.
- Storage Unit (SU) – A container for all the information on a physical file making up a component of an FU.

In the original content model (see Figure 2), the distinction between an Archival Unit and Content Unit was not well articulated. As implemented, the ConPrep system generated Content Units, which could be understood as a logical unit of content made up of one or more content files and a metadata file that captures all the relevant preservation metadata. As these Content Units were ingested into the Archive, they were renamed as “Archival Units”.

In the new content model, we refined the concepts as follows:

- Archival Unit: the abstract intellectual object
- Content Unit: a particular version (original, revision, update etc.) of the content

In effect, the presence of multiple content units within an archival unit means that the content has been sent to the archive in multiple versions by the content provider.

These versions can represent changes to the intellectual content, or technical changes such as repair of damaged files or migration to new formats by the provider. This kind of versioning is not under the control of, or initiated by, the archive, and requires maximum flexibility about the granularity and purpose (intellectual content, technical repair) of the change. In such a scenario, all versions (CUs) of an archival unit (AU) are preserved. Each version is represented by a different Content Unit, as shown in Figure 3:

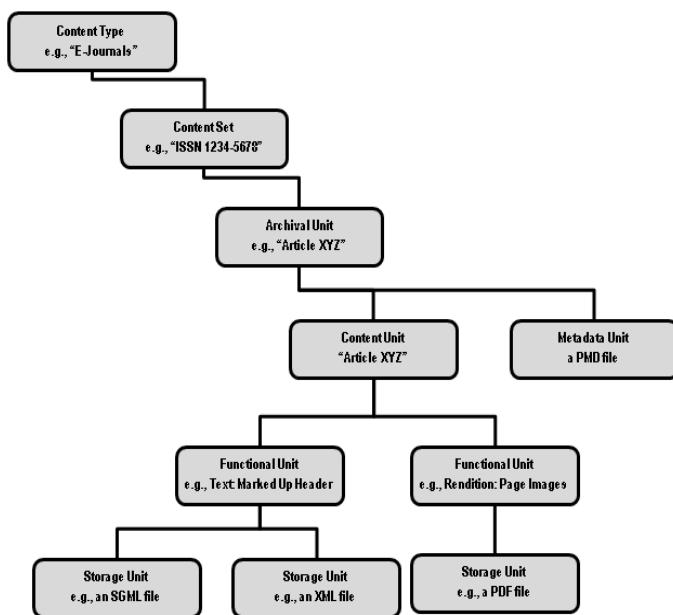


Figure 2 Portico PMETS 1.x Content Model

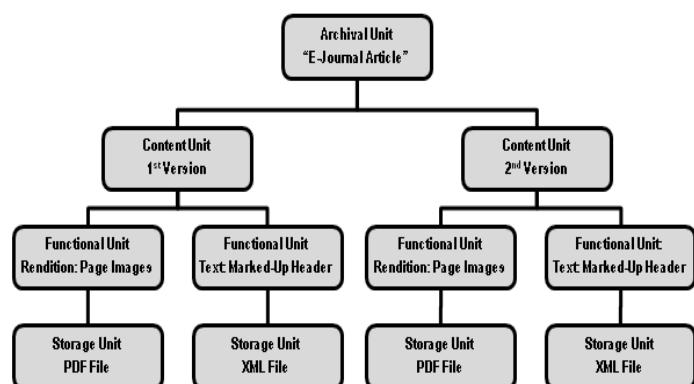


Figure 3 PMD 2.0 Content Model: Archival Unit with 2 Versions of Content Unit

In the content model, we can describe groups of Storage Units (SUs) that are "intellectually" identical but "technically" different by grouping the SUs together in one Functional unit (FU). We can use this grouping both to capture "use" information (see Figure 4), and to indicate migrated content (see Figure 5).

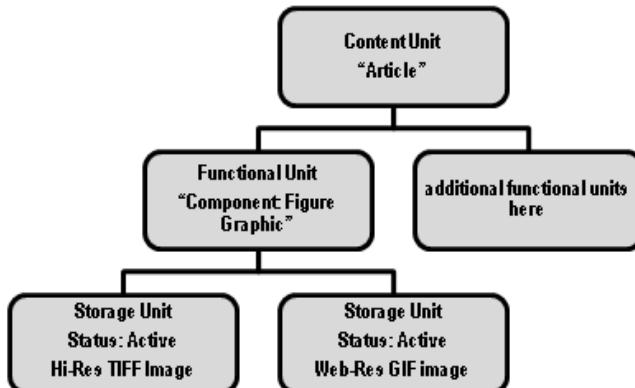


Figure 4 PMD 2.0 Content Model: Multiple Storage Units for Multiple Uses in Same Content Unit

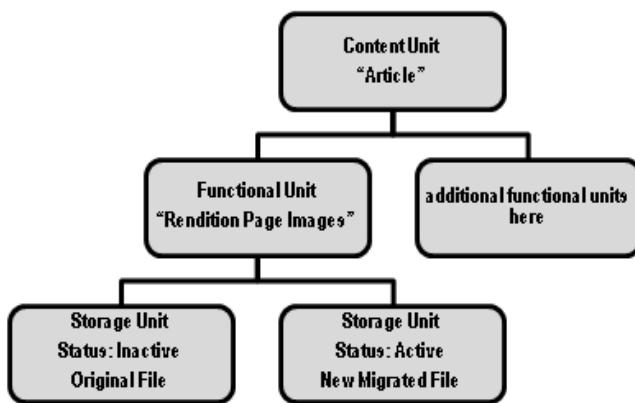


Figure 5 PMD 2.0 Content Model: Content Versioned Within Single Content Unit

Finally, in the new content model, we have extended this concept of grouping with two new components: the Storage Unit Set and the Storage Unit Pointer. These components allow us to describe, in a fairly compressed way, two new kinds of structural relationships: objects that simultaneously belong in more than one group, and relationships between sets of objects. Both are illustrated in Figure 6 below. In this example, a digitized book, each page image exists in multiple resolutions (the dotted arrows) and the entire set of high-res page images has been converted into a single PDF file (the curved red arrow). These new relationships can also be used to describe an XML text that consists of multiple files (e.g., chapters of a book).

2.1.3 Goals and Context

The goals of the new preservation metadata project were to

- Support new requirements and processes described in the previous section
- Incorporate the latest thinking from the preservation community, including from the now mature PREMIS model

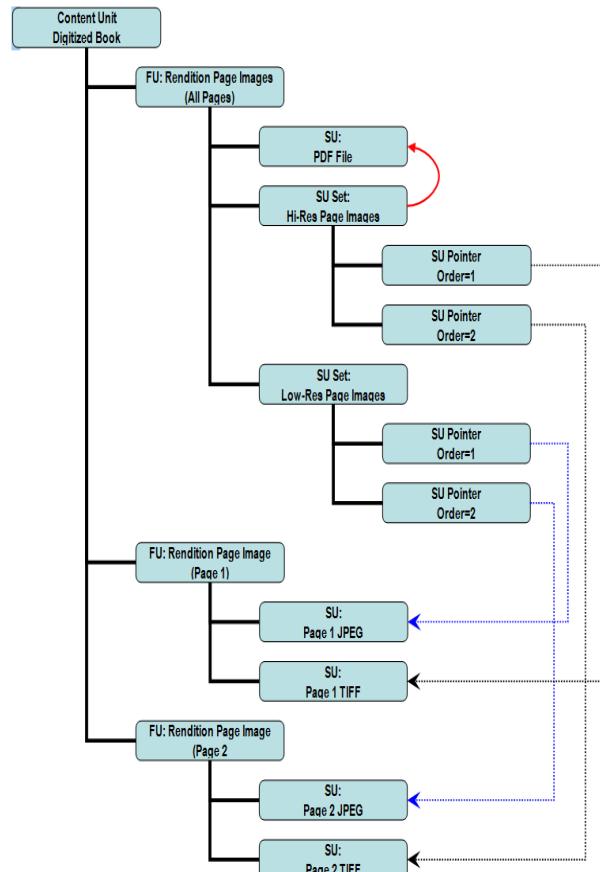


Figure 6 PMD 2.0 Content Model: Complex Component Relationships

- Develop a well-documented design for the new content model, and implement that design cleanly and consistently across all our applications. Design goals included [12]
 - Making explicit all data constraints not currently explicitly expressed in our schemas
 - Eliminating redundant information where possible
 - Establishing a clean base line for future expansion of events metadata
 - Clarifying what event goes with which object and why
 - Employing consistent editorial/coding practices (capitalization, verb tenses, etc.)

The project was undertaken as the archive continued its normal processes, including on-going incremental changes to the ConPrep system itself (deployment of new tools, facilities, etc.). It was undertaken as well in the context of a major institutional transition, as Portico, which had originally moved from a proof-

of-concept project of JSTOR to a free-standing “incubated entity” of the newly created Ithaka Harbors, in 2003, became an integrated service, along with JSTOR and Ithaka Strategy and Research, of the newly created ITHAKA, in 2009.

An additional consideration is the key role that preservation metadata plays in the archive. The archive’s preservation activities are made manifest through the preservation metadata generated and collected throughout the life cycle of a preserved object. In Portico’s case, these data can be generated during processing in ConPrep, at ingest to the archive, and as preservation activities take place thereafter.

This meant that nearly every part of the system was likely to be “touched” in some way by the metadata migration. It meant as well, as indeed Portico’s experience in scaling up ConPrep had demonstrated, that the migration would need to be carefully thought through, documented, managed, and coordinated amongst staff who would also be engaged in other work.

2.2 Planning

2.2.1 Requirements and Design: Metadata Review

Planning began with a thorough review of PMETS, including variations from version to version, and of other candidate vocabularies: METS, PREMIS, and DIDL [8].

The review of PMETS 1.x included extracting unique XPath values in actual use in PMETS files, and comparing them with possible XPaths that could be derived from the schemas, in order to determine first, if any element and/or attribute contexts proved to be unused (and possibly unnecessary), and, second, to comprehend the complete list of unique contexts and combinations of attribute/value pairs, so that all information combinations could be accommodated in a new model, and a lossless transformation accomplished.

The PMETS review enabled us to confirm an intuition of redundancy of information in each metadata file. For example, PMETS 1.x events elements included tool environment information (such as operating system and Java version in which a tool was executed). In the original design for ConPrep, we envisioned that each tool could or would run on a different server. The data model therefore provided support for capturing environment information with each individual event. However, as part of scaling up the system, we switched to embedded tool processing to gain processing efficiencies. Since almost all tools employed to process an AU are therefore run in the same environment, nearly all of the tool information in the events of a given ConPrep processing cycle will have exactly the same environment information. Additionally, we found we could flatten and simplify the structure that detailed the list of Portico and third-party tools employed in processing at each step of the workflow without loss of information.

With our new business requirements and use cases in hand, we reviewed the then current versions of METS, the PREMIS data dictionary, and DIDL. A key question to be answered was how a good a fit we could find between our requirements (and the emerging elaboration of our data model in support of them) and the expressiveness of existing, publicly available specifications. It was felt that METS was less expressive than we needed in recording the life cycle of a digital object, whether of content or of metadata. It would be difficult to record compactly the migration of individual files, or groups of files. While it was felt

to be essential to harmonize the Portico data model with key preservation information articulated in PREMIS, its data model was not entirely homomorphic with Portico’s. While the PREMIS “intellectual object” maps easily to either an AU or CU, the next level in the PREMIS model, the “representation object”, is in contrast to the Portico data model, which assumes a collection of components, some of which might constitute a complete rendition (e.g., a PDF file) of the object, and others of which might only be components from which a rendition can be created (e.g., an XML full text plus embedded images). DIDL, extended with Portico-specific attributes, looked easily extensible, but was not widely supported in a preservation context, and, with Portico attributes, would in effect be an internal format [12].

The decision was taken to develop our own schema, conformant with our data model, whose design would be optimized for the use we made of it in Java, relational database, and XML instantiations. It would be PREMIS-compliant; it could be mapped to METS; but it would be optimized for size and speed, enabling full relational normalization for use in our management database. It would make use of inheritable metadata. It would introduce a new concept: the Processing Record. This would be a block of metadata that describes all of the information common to an entire processing pass and its resulting events. One or more of these would be attached at the AU level, and could be referenced (by identifier) by subsequent objects in any CU (see Figure 7).

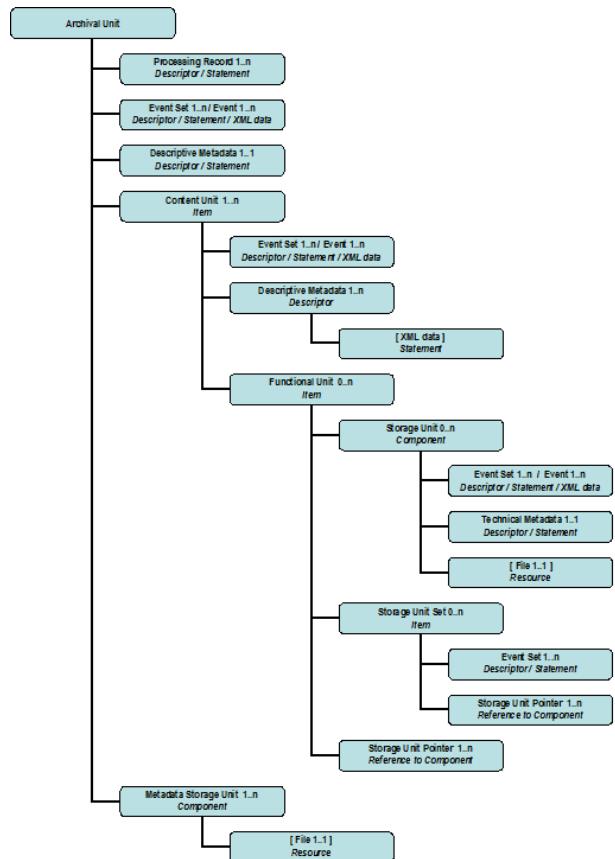


Figure 7: New Data Model: Processing Record(s) and AU, CU, and SU level Events

2.2.2 Requirements and Design: Events Review

A key component of PMETS 1.x and its underlying data model was the Portico event model. When the migration project was initiated, approximately one billion events had already been recorded in the processing of the approximately 15 million archival units and their 150 million component files. These events were associated with items in the PMETS file at both the CU and the SU level.

The event model was instantiated in the Portico Events schema. It was primarily modifications to (i.e. new versions of) the Events schema that necessitated new versions of the PMETS schema. These modifications were made incrementally, as new use cases were created by new workflow steps or other changes to the system. The event schemas defined each event separately, with different attributes and sub-elements for each event. A new design would simplify the existing data structures into a generic event that is typed with properties not specified in the schema itself, thereby allowing extensions without new versions. This in turn would obviate the need for regenerating the corresponding JAXB classes for marshalling and unmarshalling files in ConPrep.

Portico 2.0 Event Model	PREMIS Event Entity
Unique ID	eventIdentifier
Timestamp	eventDateTime
Type of Event	eventType
Rationale for the Event	eventDetail
Agent	—
User Info	linkingAgentIdentifier; linkingAgentRole
Processing Record	(not sure where to put this yet..)
Process	—
Arguments	(not sure where to put this yet..)
Input objects	linkingObjectIdentifier; linkingObjectRole
Output objects	linkingObjectIdentifier; linkingObjectRole
Tool info	(not sure where to put this yet..)
Outcome	—
Result	eventOutcome
Details	eventOutcomeDetailNote

Figure 8 Mapping New Event Model to PREMIS

We reviewed each version of the Events schema, developing tables indicating, for each activity in the ConPrep workflow, what events could result, and the element and attribute values assigned by the system. Informed by the analysis of key components of the PREMIS event model (see Figure 8), we abstracted out simple event types that describe the event itself. Those basic event types would then be qualified or sub-classed by assigning values the *Rationale* attribute. The controlled list of those values, however, would not be defined in the schema, thus allowing for extension without a new version of the schema.

2.2.3 Information Architecture

The data model having been constructed, the next steps were to review the ConPrep and archive server management Java code, and the relational database used to store and manage data object and event information during the ConPrep workflow, to determine what changes would be required to employ the new data model, and create and manage instances of the new PMD 2.0 XML format for preservation metadata. Changes included:

- New relational schema for the relational database, conforming to the new information model (see Figure 9)

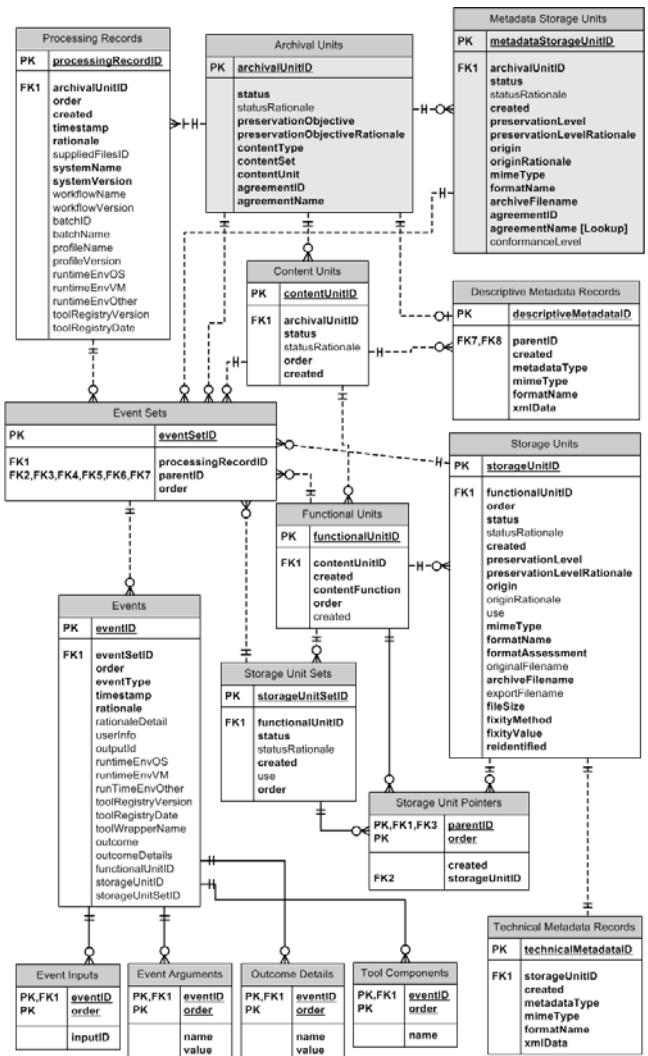


Figure 9 New Relational Schema

- New code to create, read, and write PMD 2.0 files
- New workflow step to create AU-level, Dublin Core descriptive metadata that could be employed across all content types (each CU would have content-type appropriated descriptive metadata as well)
- New code for creating instances of the new event types, with the appropriate new attribute values in managed lists, including new validator code at event creation time
- New code for the Portico delivery and audit sites for handling the new metadata files
- New tool wrapper code to employ new streamlined schema for preserving tool information
- New code for the ConPrep GUI for viewing new metadata formats, and to adapt user-defined reports to the new AU/CU hierarchy

- New Schematron validator for the new PMD 2.0 format, to enforce, among other things, controlled lists of values for event attributes
- New archive server management code to handle new PMD 2.0 format

There were other tasks associated with performing the actual migration and validation of existing PMETS files.

The first task was to create a detailed information map of the elements and attributes in the new schema (see Figure 10). This map provided a definition of the meaning of each element or attribute; its data type and constraints on values, with an indicator as to whether the constraint was to be enforced by the schema or by the Schematron validator; and its place in the relational database, in the new schema, and the corresponding element or attribute, if one existed, in the PMETS file to be migrated.

Name	Definition	Data Type and Constraints	Oracle Implementation	PMD 2.0 Implementation	PMETS 1.x Implementation
<code>xmSchemaVersion</code>	major + minor version of PMD when exported as XML	✓ fixed value of "2.0" for now; will increment in future	—	PMD @xmSchemaVersion	—
<code>conformanceVersion</code>	Name of file used to verify conformance (i.e. Schematron of this file to enumeration specifications).	String	Metadata Storage Units <code>conformanceLevel</code>	PMD @conformanceVersion	—
<code>objID</code>	Unique ID for this XML metadata file; will be the same ID as the active Metadata Storage Unit	ARK	Metadata Storage Units <code>metadataStorageUnitID</code>	PMD @objID	PorticoMETS @objID
<code>created</code>	Date/time this record (metadata file) was created; also appears on the active Metadata Storage Unit	✓ Timestamp	Metadata Storage Units <code>created</code>	PMD @created	PorticoMETS <code>metaObjId</code> @CREATEDATE
ARCHIVAL UNIT					
<code>archivalUnit</code>	The basic unit of archived content.	✓ Children 1..n Processing Record 1..n Event Set 1..n Descriptive Metadata 1..n Content Unit 1..n Metadata Storage	Archival Units	PMD @ArchivalUnit	PorticoMETS

Figure 10 Information Mapping

The next task was to develop the transformation and validation pipeline for the existing 15 million PMETS files. This entailed

- Extracting a copy of the files from the Portico archive
- Developing an XSL transformation from PMETS to PMD, using the information mapping table
- Developing the Schematron assertions to test the data types and constraints in the information mapping table (this is the same Schematron that would be used in ConPrep, going forward, to validate new PMD files)

The pipeline was to be run via an application called “ConprepLite.” ConprepLite is a light-weight façade over the Conprep workflow and tool wrapper classes. It was devised to enable the Portico Data Team to test their transformation and validation tools against thousands of files, while using the same code invoked by the ConPrep runtime to run those tools. Because we were scaling up the use of ConprepLite from thousands to millions of files, it was also necessary to refactor the ConprepLite software to be multithreaded, and to streamline the reading and processing of the XML configuration files (which listed input files, and the workflow steps to be executed) from a document object model to a streaming model.

Finally, we would require one set of scripts to extract samples of the newly created AU-level descriptive metadata, for review and approval by the Portico Archive Service Product Manager, and another set of scripts to import the new PMD 2.0 files into the archive, and update the archive management database to reflect the presence of these new assets, and their relationship to the existing content and metadata files.

2.3 Execution

2.3.1 Technical Challenges

One of the lessons learned from scaling up the ConPrep system was to “expect surprises” [11]. That expectation was amply met when we revved up the pipeline. We found that processing such a large number of (often very) large XML files stressed both hardware and almost every layer of software in the pipeline stack.

Tuning of all sorts was an issue. With multiple threads running on multiple machines, it took some tuning to settle on reasonable batch sizes, so that any failure of a single batch would not result in the waste of days or even weeks of run time. It took some trials to determine the optimum thread count to employ on each instance of ConprepLite that was running on multiple, and different, hardware and operating systems configurations.

Both the PMETS files and the XSL files designed to transform them were quite large and complex (the transform files run to approximately 3000 lines of code). The PMETS files also contained segments from many different namespaces: the PMETS namespaces, Dublin Core, three namespaces in the JHOVE technical metadata, and so on. These namespaces appear scattered throughout the XML document tree, which could often be quite deep. At times, this broke the name pool limit in the version of the Saxon XSL transform engine we were using. We had to upgrade and test our transform with a later version. Additionally, even with the newer version, files with very deep technical metadata trees resulted in stack overflow. We had to tune our memory allocation to handle this (eventually ending up with a 30 gigabyte heap size).

Handling large-scale numbers of very large files resulted in many different kinds of memory tuning. Having moved first from 32-bit to 64-bit Java Virtual Machines (JVM), we found it necessary to increase the JVM permgen space in setting the JVM environment at run time. We then found we had to tune the size of the pool allocated for interned strings, as we were overrunning standard limits for that as well.

ConprepLite creates many directories and files as intermediate artifacts of conversion and validation. Some of the ConprepLite instances were running on machines with older versions of UNIX. These instances ran into difficulties when the number of directories exceeded the maximum limit for child inodes on these systems.

Part of the PMETS-to-PMD2 transformation included the creation of an Archival Resource Key [9], used as an object identifier for nearly every element in the schema. We found that the NOID minter was not able to keep up with the number of requests being made by multiple ConprepLite instances. We established a separate NOID minter server per process to handle this.

The ConprepLite pipeline consisted of three steps: transformation from PMETS to PMD2, validation of the PMD2 file against the PMD2 schema, and further validation of the PMD2 file with Schematron. The pipeline was running quite slowly at first. We

looked to see if it was IO-bound or process-bound. It turned out to be the latter, with resources being consumed largely by the user rather than the kernel. The ConprepLite instances were then moved to heavier-duty machines with an NFS mount to the file system with the extracted PMETS files.

Additionally, inspecting the logs, we saw that nearly two-thirds of the time was being spent on the Schematron validation. Our first thought was that the heavy use of regular expressions was consuming a lot of processing time. This however proved not to be the case. We then recollected that Schematron essentially is a code generator, taking as input user assertions, and transforming them against a “skeleton” to generate an XSL transform actually run against the file being validated. We had already optimized Conprep and ConprepLite to cache compiled XSL transformations, including the XSL transform generated “on the fly” by Schematron the first time it is invoked in the workflow. Outside the ConprepLite workflow, we serialized the XSL transform generated by Schematron, so that we could inspect the generated code to see what actually was being run. What we found was that Schematron’s generated code was using a technique (XSL “modes”) which resulted in over 128 passes through each of the (very large) PMD2 files. We tuned the code to minimize passes through the PMD2 files.

2.3.2 Quality Assurance

Although the transformation was tested against many sample files as it was developed, we expected to encounter, in a transformation of such complexity, dealing with input of such complexity, errors of one sort or another, as we in fact did. Key to catching such errors was the capability for large-scale automated validation, both via schema validation and Schematron.

We also performed extracts of the newly generated descriptive metadata for manual review, to verify the correctness of the newly created metadata.

As a matter of policy, Portico retains the original PMETS file along with the new PMD file (which references the now-inactive earlier version) associated with the archival unit. This enables us to re-run the transform as needed, should we discover, at a later time, any errors in our transformation process.

3. REFLECTIONS

It is important to consider the process of migration, not just from the perspective of issues raised by specific file formats, but also in the larger context of the life cycles of systems and software themselves, and in the new use cases for repository content that emerge from ever-evolving expectations of an archive’s community of use. As Portico’s experience with its preservation metadata would seem to indicate, it is reasonable to expect over the long term that changes in the large-scale infrastructure of archives, their capabilities, and their communities of use, will themselves necessitate the ability to manage, manipulate, move, and migrate content at very large scales.

Archives and repositories will need to make their own assessments of the necessity, feasibility, and usefulness of such large-scale asset migrations as Portico undertook. They will need to balance the tradeoffs between just-in-time versus large scale pre-emptive migration. And they will need to make these assessments not only about both assets conventionally understood as “content”, but about system-generated artifacts such as

preservation metadata, which also constitute content, albeit of a less conventional kind, in need of stewardship and preservation.

Preservation institutions will need to assess the likely “lossiness” of such migrations. It is comparatively easy to determine the significant properties [6] to be tracked in an XML-to-XML migration such as the one described in this paper. Nevertheless, it is important to articulate that mapping in advance of the transformation, so that the success of the transformation can be tested. This is crucial for the construction of automated tests of the correctness of the transformation – another key capability for migration at scale.

Fifteen million of anything is a lot. It is no surprise that it takes a lot of work to manipulate content at that scale, whether that manipulation is a migration, or some other operation. In this case, in terms of elapsed time, Portico spent approximately three to four months planning the migration, and another nine months in its development and execution.

Given the scale at which this was happening, the importance of the content itself, and the many other activities of the staff involved in accomplishing a migration or any similar large-scale, cross-corpus manipulation of content, it is crucially important carefully to analyze, document, plan, and track such efforts. An important part of the planning will be to expect – and to allow time and resources for --the unexpected.

4. ACKNOWLEDGEMENTS

The authors would like to acknowledge Evan P. Owens, formerly CTO of Portico, and Vice President for Content Management, ITHAKA, who directed the migration project, the project documents of which were key source materials for this paper.

5. REFERENCES

- [1] Beck, Jeff. Report from the Field: PubMed Central, an XML-based Archive of Life Sciences Journal Articles. Presented at International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML, Montréal, Canada, August 2, 2010. In *Proceedings of the International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML*. Balisage Series on Markup Technologies, vol. 6 (2010). DOI=10.4242/BalisageVol6.Beck01.
- [2] Caplan, Priscilla. The Florida Digital Archive and DAITSS: a Working Preservation Repository Based on Format Migration. *International Journal on Digital Libraries* 6.4 (2007): 305–311.
- [3] CCSDS. *Reference Model for an Open Archival Information System (OAIS)*. CCDS 650.0-B-1 Blue Book Issue 1 (2002)
- [4] Digital Library Federation. Metadata Encoding and Transmission Standard (METS) Version 1.7. 2008 Web 06 June 2012 from <http://www.loc.gov/standards/mets/version17/mets.xsd>
- [5] Heslop, H., Davis, S. & Wilson, A. *An approach to the preservation of digital records* (2002) Web 08 June, 2012, from http://web.archive.org/web/20031217152126/http://www.naa.gov.au/recordkeeping/er/digital_preservation/Green_Paper.pdf

- [6] Hedstrom, M., and C. A. Lee. Significant Properties of Digital Objects: Definitions, Applications, Implications. *Proceedings of the DLM-Forum*. 2002.
- [7] ISO/IEC 19757-3:2006 Information technology -- Document Schema Definition Language (DSDL) -- Part 3: Rule-based validation – Schematron ISO/IEC 2006
- [8] Declaration ISO/IEC JTC 1/SC 29 N 3971 Information Technology — Multimedia Framework — Part 2: Digital Item
- [9] Kunze, J. and Rodgers, R. *The ARK Identifier Scheme*. 22 May 2008. Web 06 June 2012 from <https://confluence.ucop.edu/download/attachments/16744455/arkspec.pdf?version=1&modificationDate=1261036800000>
- [10] PREMIS Editorial Committee. PREMIS Data Dictionary, Version 2. Library of Congress March 2008 Web 06 June 2012 from <http://www.loc.gov/standards/premis/v2/premis-dd-2-0.pdf>
- [11] Owens, Evan, Cheruku , Vinay, Meyer, John, and Morrissey, Sheila. Digital Content Management at Scale: A Case Study from Portico. Presented at *DLF Spring Forum*, Minneapolis, April 28-30, 2008. Web 06 June 2012 from <http://www.diglib.org/forums/spring2008/presentations/Owens.pdf>
- [12] Owens, Evan. ITHAKA Preservation Metadata 2.0: Revising the Event Model. Presented at PREMIS Implementation Fair 2009. Web 06 June 2012 from <http://www.loc.gov/standards/premis/pif-presentations/Portico PREMIS Workshop.ppt>
- [13] Van Wijk, Caroline. “KB and Migration Test Plan”. National Library of the Netherlands (KB), Digital Preservation Department. 6 November 2006. Web 29 May 2012, from http://www.kb.nl/hrd/dd/dd_projecten/KB%20and%20Migration%20Test%20Plan.pdf

Managing multidisciplinary research data

Extending DSpace to enable long-term preservation of tabular datasets

João Rocha da Silva *
INESC TEC / DEI, Faculdade
de Engenharia, Universidade
do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto PORTUGAL
joaorosilva@gmail.com

Cristina Ribeiro
INESC TEC / DEI, Faculdade
de Engenharia, Universidade
do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto PORTUGAL
mcr@fe.up.pt

João Correia Lopes
INESC TEC / DEI, Faculdade
de Engenharia, Universidade
do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto PORTUGAL
jlopes@fe.up.pt

ABSTRACT

In a recent scoping study we have inquired into the data management needs of several research groups at the University of Porto and concluded that data quality and ease of on-line data manipulation are among the most valued features of a data repository. This paper describes the ensuing approach to data curation, designed to streamline the data depositing process and built on two components: a curation workflow and a data repository. The workflow involves a data curator who will assist researchers in providing meaningful descriptions for their data, while a DSpace repository was customised to satisfy common data deposit and exploration requirements. Storing the datasets as XML documents, the repository allows curators to deposit new datasets using Excel spreadsheets as an intermediate format, allowing the data to be queried on-line and the results retrieved in the same format. This dedicated repository provides the grounds for collecting researcher feedback on the curation process.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Digital Libraries; H.4 [Information Systems Applications]: Miscellaneous

General Terms

Human Factors, Management, Standardisation

Keywords

Research data management, data repositories, DSpace extension, digital curation

*Supported by research grant SFRH/BD/77092/2011, provided by the FCT (Fundação para a Ciência e Tecnologia).

1. INTRODUCTION

The need to adopt effective data management procedures as part of the research workflow is currently assuming great importance, with data management requirements being imposed by research funding institutions. An example is the NSF, which now requires the inclusion of a data management plan in every research grant proposal [1]. In the UK, JISC has recently launched the Managing Research Data programme covering aspects such as infrastructures, data management plans and supporting technologies [3]; also, the Digital Curation Centre¹ provides resources and consultancy for researchers. Besides official policies, researchers are also becoming aware of the scientific impact of their data assets [5]. Universities are realising the benefits of exposing their research in institutional repositories and are seeking to extend them to research data. Several scoping studies and projects such as the DAF (Data Asset Framework) [9], the Edinburgh DataShare [11] or the DANS (Data Archiving and Networked Services) [2] have yielded data management workflows and recommendations. The management of research data requires a deep involvement of the researchers, since they are both creators and consumers of datasets. Thus, they must be involved in the dataset preparation and description process necessary to make the data available in the repository. Datasets pose hard problems regarding preservation formats, an issue that has been addressed by the MIXED project with a rich XML schema that can be used for the preservation of Excel spreadsheets through intermediate XML formats [12]. More general solutions for format identification and validation include the JHOVE² and the DROID³. After learning from the conclusions of large projects in this field, we designed our repository as a tool to support the researcher throughout the research workflow, continuing a previous work [8] in which we have detailed the architecture of a data management repository for U.Porto (University of Porto). Since then, a set of open-source modules have been combined with the DSpace platform, yielding a prototype that can be used by researchers from different domains for recording, sharing and preserving tabular data, allowing us to gather additional researcher feedback.

¹<http://www.dcc.ac.uk/>

²JSTOR/Harvard Object Validation Environment : <http://hul.harvard.edu/jhove/>

³Digital Record Object Identification : <http://sourceforge.net/apps/mediawiki/droid>

2. SCOPING STUDY AND DEVELOPMENT CONTEXT

In order for research data repositories to become an integral part of the research workflow they must provide an added value to researchers throughout their research in return for their assistance in the curation of the datasets that they produce. With this in mind, we have designed a data curation workflow to handle the data management needs of active research groups at U.Porto. It is recognised that data curation should not be performed only at the end of the research workflow but rather follow the research process and start as early as the raw data is gathered from sensors or other equipment [4]. Later on, as researchers fully understand the concepts necessary to describe their domain of study [10], detailed annotation should be arranged between researchers and curators, quickly enough to make it possible for the researcher to cite the datasets in a publication. After the work is published, it should be made available at a publication repository with a direct link to the data—the latter being stored in a specialised repository offering data exploration and annotation features. This workflow was designed starting from the experience of previous work in this field [2, 9, 11] combined with our results from a scoping study involving several research groups at U.Porto[6]. Researchers interviewed in our study often stated that the added value of a repository depends on the ability to query and explore parts of the deposited data using their web browser, and also on the existence of metadata comprehensive and accurate enough to make it possible for them to reuse the data. At the organisational level, this means that research activities should be supported by a curation service that helps researchers maintain their data—a process that must begin early in the research process. From an engineering point of view, this poses several challenges that we have started to study at U.Porto by taking a standard DSpace repository and extending the underlying data model to allow for finer-granularity data access and metadata annotation, while maintaining the user-friendliness of its user interface.

3. DATA ACCESS AND PRESERVATION

Implementing a data repository using DSpace proved to be a challenge because its underlying data model is not designed to handle querying and exploring tabular datasets at table row granularity. In the standard DSpace data model, the smallest-granularity entity to which metadata can be added is an **Item**, which groups a set of data files representing the authors' work. This does not allow the system to retrieve parts of the data inside a file, requiring the user to download it as a whole and then explore the data using the program that the researcher originally used to create and manipulate it. This dependency on the original software used to prepare the data is often a reason for data loss as that software may become obsolete. The curation process starts with the standard DSpace workflow for self-deposit, after which an **Item** containing all the data files pertaining to the new dataset is created. Curators may then access a curation page (that we implemented) to upload custom-defined Excel workbooks [8] containing the tabular data originally present in the **Item**'s files, associating those tables with the files. When an Excel workbook is uploaded, the repository translates the data into an XML document and stores it in the DSpace database. This provides DSpace with the flexibility

to have as many columns in a data table as necessary, regardless of their datatypes (**integer**, **string**...), something not originally possible given DSpace's relational model. XML documents are also easy to query using XQuery and are much more suitable for long-term preservation than their original counterparts. After a file is curated in this way, it becomes accessible to all registered users through a data exploration tool that allows users to restrict the visible rows by applying various filters on the columns directly from the Web browser. It is also possible to download the filtered data as an Excel workbook—the Excel format is only used as an intermediate format, not as the core storage of the repository.

<http://www.w3.org/2001/XMLSchema>

3.1 An XML format for tabular data

The structure of the XML representation of the data tables is presented in a systematic manner in Figure 1. The XML schema for the structure is located at <http://www.xml-cml.org/schema/schema3/schema.xsd>.

```
<?xml version="1.0" encoding="UTF-8"?>
<?xsi:include href="http://dublincore.org/schemas/xmls/adc/dc.xsd"?>
<?xsi:include href="http://www.w3.org/2001/XMLSchema"?>
<?xsi:include href="http://www.xml-cml.org/schema/schema3/schema.xsd"?>

<Element: tables
  Type: complex
  [1..1] sequence of {
    <Element: table
      Type: complex
      [1..1] sequence of {
        <Element: record
          Type: complex
          [1..1] sequence of {
            <Element: metadata
              Type: complex
              [1..1] sequence of {
                <Element: choice of {
                  <Element: Group + dc:elementsGroup
                  <Element: + cml:formula
                }
              }
            }
          }
        <Element: data
          Type: complex
          [1..1] sequence of {
            <Element: rows
              Type: complex
              [1..1] sequence of {
                <Element: row
                  Type: complex
                  [1..1] sequence of {
                    <Element: choice of {
                      <Element: + cml:formula
                    }
                  }
                }
              }
            }
          }
        <Element: headers
          Type: complex
          [1..1] sequence of {
            <Element: header as string
          }
        }
      }
    >
  >
Attribute: index as xs:integer
```

1

Figure 1: An example of the XML documents stored in DSpace, containing tabular data and their metadata

The structure of the documents includes a root element called **tables**, and contains a series of **table** elements. Each **table** contains a **metadata** section delimiting a sequence of qualified elements and their respective values. We need to incorporate elements from different XML schemas depending on the domain of the research dataset, so the **metadata** section can include elements from the Dublin Core schema as well as others from different metadata profiles, as is illustrated by the inclusion of the **cml:formula** element from the CML⁴ schema. The **headers** section contains a list of all the table headers for this table (qualified metadata elements from arbitrary schemas), and the **data** section contains a series of **rows** from the current **table**. Each **row** contains a series of cells that match qualified elements and corresponding

⁴Chemical Markup Language <http://www.xml-cml.org/>

Figure 2: Dynamic grid interface used to explore datasets

Figure 3: Search interface used for retrieving selected data tables

values. The format does not currently differentiate string, integer or any other datatypes for the contents of each cell, like existing profiles such as MIXED by DANS[12]. The MIXED project has proposed a richer XML schema for tabular data; we have chosen a lighter XML model for our data, for the sake of building a complete workflow where we can provide a fast prototype and use it to evaluate with our users the satisfaction of their requirements. Since the visualisation component relies on XML Stylesheets to convert the documents stored in the database to the format accepted by jqGrid (a jQuery component used for data presentation) it is possible to make changes to the internal schema without major changes in the code.

4. A WALKTHROUGH OF THE SOLUTION

The data curation workflow begins with a meeting between the researcher and the curator. In this meeting, the dataset that the researcher wishes to deposit goes through the standard DSpace depositing workflow, and a new Item is created. This Item will have its own metadata and all the files created by the researcher, exactly as they were originally produced. The next stage in our proposed workflow is the intermediate curation step for tabular data, in which the curator accesses the newly created DSpace Item's page to retrieve each file and build an Excel workbook containing the data inside it, as well as any relevant metadata—each of the sheets of this Excel workbook contains a single table, as well as matching

table-level metadata. A workbook is built for each file and sent to the repository through a specific link (see area 3 of Figure 4), after which it is parsed by the system and translated into an XML document that is stored in the DSpace database. The Excel spreadsheet is discarded since it is only the intermediate format. After a file is curated, its tables can be explored through the data navigation interface (shown in Figure 2) and retrieved by specifying some of their columns on the querying interface (shown in Figure 3).

This interface is accessible from the Item page through the “Explore Data” link that we have added (Figure 4, area 2), next to the option that is normally used to download the file (area 1). When the “Explore Data” option is selected, the tables contained in the file are shown in a dynamic grid—shown in Figure 2—allowing the user to filter the data directly from the Web browser. In Figure 2, button 1 allows users to specify combinations of restrictions on each of the table’s columns (which will appear in area 2). The user may add more restrictions by selecting button 3 or execute the filtering by selecting button 4. At any time, the user may download the selected data, the currently selected table or the whole workbook with all tables in the file. Both the data and metadata are provided in Excel format when the user selects the desired option from area 5. These use cases can be seen at the project’s documentation wiki [6], including several videos [7] designed to demonstrate how curators and researchers can interact with the developed system.

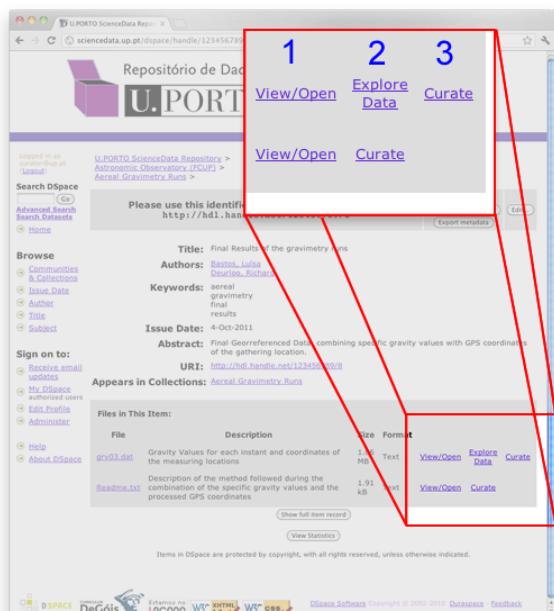


Figure 4: Extra “Curate” and “Explore” options were added to the DSpace Item exploration page

5. CONCLUSIONS AND FUTURE WORK

We have successfully implemented a curation workflow based on the needs of several U.Porto research groups. It uses an extended DSpace instance as the supporting platform, taking advantage of its effective built-in workflow engine for the self-deposit of datasets. The implemented extensions provide additional curation features designed to access the data at a fine granularity. The end result is a data exploration

interface that allows users to query the data directly from their Web browsers and, if they wish, download the results in Excel format. The core data storage uses an XML format for enhanced long-term preservation, while Excel workbooks are used as vehicles for data transfer, therefore improving the user-friendliness of the system. The “look-and-feel” that users are accustomed to finding in a DSpace repository was maintained in the interface design, in an effort to provide a consistent user experience throughout the whole extended platform.

We are now in the process of validating the prototype through a second round of interviews with the researchers that participated in the scoping study. As for the continuation of the developments, we are now focusing on improving the methods through which researchers can retrieve datasets (by extending the DSpace free search capability to index the contents of our research datasets) and also in finding similarities between them using their metadata.

6. REFERENCES

- [1] N. S. Foundation. NSF Data Management Plan Requirements. <http://www.nsf.gov/pubs/policydocs/grantsgovguide0111.pdf>, January 2011.
- [2] Ingrid Dillo and Peter Doorn. *The Dutch data landscape in 32 interviews and a survey*. 2001.
- [3] JISC. Research Data Management Infrastructure Projects, 2011.
- [4] L. Lyon. Dealing with Data: Roles, Rights, Responsibilities and Relationships, 2007.
- [5] H. A. Piwowar, R. S. Day, and D. B. Fridsma. Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, 2(3):e308, 03 2007.
- [6] J. Rocha da Silva, C. Ribeiro, and J. Correia Lopes. UPData - Scientific Data Curation at U.Porto. <http://joaorosilva.no-ip.org/updata/wiki/doku.php>.
- [7] J. Rocha da Silva, C. Ribeiro, and J. Correia Lopes. UPData - Scientific Data Curation at U.Porto - Demonstration Videos. http://sciencedata.up.pt/doc/doku.php?id=demo_videos.
- [8] J. Rocha da Silva, C. Ribeiro, and J. Correia Lopes. UPData A Data Curation Experiment at U.Porto using DSpace. In *iPRES 2011 Proceedings*, pages 224–227, 2011.
- [9] Sarah Jones. Data Audit Framework lessons learned report: GUARD Audit.
- [10] A. Tonge and P. Morgan. SPECTRa - Submission, Preservation and Exposure of Chemistry Teaching and Research Data. Technical report, Cambridge University, Imperial College (JISC Digital Repositories Programme), 2007. <http://lib.cam.ac.uk/spectra/FinalReport.html>.
- [11] University of Edinburgh. What is Edinburgh Datashare? <http://datashare.is.ed.ac.uk/>.
- [12] R. van Horik and D. Roorda. MIXED: Repository of Durable File Format Conversions. In *iPRES 2009 Proceedings*, 2009.

On the Applicability of Workflow Management Systems for the Preservation of Business Processes

Rudolf Mayer
Secure Business Austria
Vienna, Austria
rmayer@sba-research.at

Stefan Proell
Secure Business Austria
Vienna, Austria
sproell@sba-research.at

Andreas Rauber
Secure Business Austria
Vienna, Austria
arauber@sba-research.at

ABSTRACT

Digital preservation research has increasingly been shifting focus from the preservation of data and static objects to investigate the preservation of complete processes and workflows. Capturing all aspects of a process to be preserved, however, is an extensive and difficult undertaking, as it requires capturing complete software setups with potentially complex setups. Further, the process might use external services that are not easy to capture and monitor. In this paper, we therefore investigate the applicability and usability of using Workflow Management Systems for executing processes in a standard environment where the state of the process can be closely monitored. To this end, we compare three popular workflow management systems. We use a scenario from eScience, implementing a data analysis process, to evaluate the challenges and implications for establishing sustainable and verifiable eScience processes.

General Terms

E-Science, Research Infrastructures, Process Preservation

1. INTRODUCTION

Classical digital preservation has its focus on maintaining the accessibility of digital objects, such as documents, images or other rather static digital files. This area is well understood and solutions to many former problems exist. The next stage of preservation research deals with the preservation of complex systems and composite processes, as they can be often found in the business and research community. Such processes are in contrast to static files highly dynamic and require constant monitoring. The process orientation can be identified in rather young research areas such as E-Science and also the discipline of business process engineering. In science, experiments need to be preserved as researchers need to be able to reproduce and build on top of earlier experiments to verify and expand on the results. It may also prove essential to understand any pre-processing steps and consequences on the interpretation of results in

any future meta-studies building on top of earlier research results. In businesses, preservation of processes can play an important role e.g. in liability cases, where a company has to prove that a certain series of steps was executed in the correct manner and according to standards, best practices or laws and regulations. Another motivation are patent litigations, when a company would want to demonstrate how a certain invention originated. Therefore, businesses have to preserve their processes for many years and need to rerun them whenever necessary.

Both areas have in common that they involve large amounts of data and integrate heterogeneous services. These systems form critical infrastructure. Hence their preservation is an urgent and important quest that needs to be tackled. This is a challenging task as these systems are highly complex and consist of many different components, not all of which are under the influence of one controlling instance.

Processes describe how a certain goal has to be achieved. Scientific and business processes consist of intermediate steps which are modelled by the use of workflows. There exist workflow management systems for both domains. These are generic software systems that are driven by explicit process designs to enact and manage operational business or scientific processes, as defined by Aalst[13]. A workflow is defined as “The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules.”[13]. Hence, workflows describe the flow of information through a business process. The same paradigm has been adapted in the scientific domain, which lead to Scientific Workflow Management Systems that aid scientists to handle increasingly complex and data driven experiments. In order to tackle this increasing complexity and the orchestration of manifold services and systems, the concept of scientific workflows has received increasing attention within the research community. E-Science projects profit from the combination of automated processing steps in workflows in order to perform complex calculations and data transformations. The advantage of workflows is their capability of adding structure to a series of tasks. They can be visualized as graph representations, where nodes denote processes or tasks and edges denote information or data flows between the tasks. This adds a layer of abstraction and helps to clarify interactions between tasks [2].

Many of today’s data-intensive experiments depend on a

number of external service such as Web services, or continuously changing third-party libraries and applications. These changes are not always under the control of the researcher, and may happen at a system level beyond the awareness of the individual researcher, such as e.g. a new library being installed as part of (automatic) system maintenance. This may lead to different results from the workflow, or render the workflow not executable altogether. The possibility to reproduce workflows is also a crucial principle in the business domain.

Preserving the repeatability of such a process in a changing technological environment is therefore a current and emerging topic in Digital Preservation research. Digital preservation of business or E-Science processes requires capturing the whole context of the process, including e.g. dependencies on other computing systems, the data consumed and generated, and more high-level information such as the goals of the process. In this paper, we investigate the feasibility of Workflow Management Systems (WFMS) for preserving scientific processes. We propose that the implementation of a scientific process can be seen as migration strategy, as the original design, structure, meaning and results can be preserved. We provide an overview on the power of such systems and evaluate the effort to migrate workflows between different WFMSs.

The success of preservation activities have to be evaluated. Hence it is required to identify and examine all involved components and the data exchanged between them. This can be achieved by the use of provenance data, which describe the lineage of data and the causal relationships between intermediate steps. Most WFMSs provide the possibility to generate such provenance data automatically. Therefore these systems are valuable for the preservation of processes. In this paper, we first outline which information is important to be captured. We will then investigate the suitability of such automatically recorded provenance data in a case-study of a scientific experiment in the data mining domain.

2. WORKFLOWS AND WORKFLOW MANAGEMENT SYSTEMS

In both domains - science and business - workflows allow to precisely define the involved steps, the required context and the data flow between components. The modelling of workflows can be seen as an abstraction layer, as they describe the computational ecosystem of the software used during a process. Additionally, they provide an execution environment, that integrates the required components for performing a process and executing all defined subtasks. This abstraction supports the preservation process as there is more information about the execution details available. Hence we examine the feasibility of scientific workflow systems for the preservation of scientific processes.

Different scientific workflow management systems (SWMS) exist that allow scientists to combine services and infrastructure for their research. The most prominent examples of such systems are Taverna [9] and Kepler [5]. Vistrails [11] is another workflow management system prominent especially in visualisation, but will not be covered in detail here. Workflow management systems are also prominently used to execute business processes. We will look at the open-source

system Activiti.

2.1 Taverna Workbench

Taverna Workbench¹ is an open source project that allows to design and run workflows. It is a general purpose workflow engine that can be used for various applications. It is written in the Java programming language and distributed under the GNU Lesser General Public License (LGPL²).

Taverna allows to orchestrate various services and to model the data flow between its components in order to automate a process. Therefore Taverna is widely used in the scientific community and used for modelling data centric experiments. It provides a graphical user interface that allows scientists to design and execute their experiments in a convenient way and to visualize the data flow of an experiment. An example of such a workflow is given in figure 1.

Taverna is a service oriented workflow engine and allows to solve tasks by using either local or remote services. Local services include basic file operations, format conversions and many different tools. Remote services include predefined Web services from various domains, such as bioinformatics or chemistry. It is also possible to implement custom services using the Taverna Java Application Programming Interface (API). Services can also be implemented via scripting language; Taverna to this end supports the language *beanshell*, which is based on the Java programming language.

Taverna uses ports to exchange data between the services: each service can have several input and output ports, where one output port serves as input for a subsequent service. The workbench has an integrated support for scalars and lists, which includes implicit iteration over arrays of data. Looping over data elements is also integrated and allows the usage of control and synchronization points. In its basic configuration, Taverna simply passes down data tokens in a downstream fashion to the next connected service.

2.2 The Kepler Project

The Kepler scientific workflow system[6] is a general purpose application suite for managing, orchestrating and executing scientific workflows. Kepler is an open source project, distributed under BSD license³ and written in the Java programming language. It provides a graphical interface which enables scientists to design and execute experiments, by linking various services each fulfilling a subtask of a workflow. Kepler inherited the graphical user interface and the actor-centric view of workflows from the Ptolemy II⁴ project, which is a framework for designing embedded systems and the communication between components.

Actors are used to model individual steps during the execution of an experiment; they can perform relatively simple tasks as format conversions, displaying data or reading a file from a Web server. There also exist more complex actors that invoke specialized grid services, utilize domain specific databases or execute external services. It is also possible

¹www.taverna.org.uk/

²www.gnu.org/licenses/lgpl.html

³<http://www.opensource.org/licenses/bsd-license.php>

⁴<http://ptolemy.eecs.berkeley.edu/ptolemyII/>

to develop custom actors by implementing desired features in Java using the Kepler API, and instantiate them within the Kepler system. Further more, python scripts can be executed as well, which reduces the development effort and enhances the flexibility, as no detailed knowledge about the Kepler API is needed.

Actors use Ports for communicating and exchanging data with each other; each Port can either serve as input, output or both to an actor. Ports connect Actors by using channels, which models the data flow and logical sequence of intermediate steps within the workflow. Actors are therefore roughly comparable to services in Taverna.

The workflow is orchestrated by a so-called director, which is the component responsible for arranging the timing of the data flow. There exist different directors for various purposes, such as sequential, dynamic or parallel execution of actors.

2.3 Activiti

Activiti is a workflow and Business Process Management (BPM) Platform, based on the Business Process Modelling Notation (BPMN) 2.0. It is available as open source software and written in the Java programming language, maintained by a consortium of companies offering cloud and Java solutions.

Unlike Taverna or Kepler, it doesn't provide an integrated GUI. Instead, the design of the workflows is enabled by an BPMN 2.0 editor which can be installed as an extension to the Eclipse Integrated development environment (IDE). All the elements available in BPMN 2.0 can thus be used to design the workflow. For execution of the workflow, Activiti can be run as a web-application on a Java Application Server, or as a stand-alone Java application.

Of the BPMN 2.0 elements, most importantly, tasks represent processing steps in the workflow. These tasks are associated via simple sequence flow connections to define the order of execution. Control flow can be modelled with gateways, such as for parallel or exclusive processing. There is no explicit definition of data exchanged, as it is done via Ports in Taverna or Kepler. Rather, a global state of data variables is kept in a key-value map.

Implementation of tasks is enabled by Java classes, or scripting languages that support the Java Scripting Platform, which includes among others JavaScript, Python, Ruby, and Groovy. Both are straight-forward with convenient integration into the BPMN editor. User interaction tasks, which play a more important role in business processes than in scientific experiments, can be implemented via forms; these enable the user to input data, e.g. as workflow input parameters. Unlike the scientific workflow management systems of Taverna and Kepler, Activiti doesn't provide a library of pre-defined tasks to use; however, in the implementation one can draw on the many libraries available to Java and all the script languages supported.

3. CASE STUDY - SCIENTIFIC DATA MINING PROCESS

In this section we discuss the implementation of a typical E-Science process with the workflow management systems introduced above. The specific process used in our case study is a scientific experiment in the domain of data mining, where the researcher performs an automatic classification of music into a set of predefined categories. This type of experiment is a standard scenario in music information retrieval research, and is used with many slight variations in set-up for numerous evaluation settings, ranging from ad-hoc experiments to benchmark evaluations such as e.g. the MIREX genre classification or artist identification tasks [7].

The experiment involves several steps, which can partially be parallelised. First, music data is acquired from sources such as benchmark repositories or, in more complex settings, online content providers, and in the same time, genre assignments for the pieces of music are obtained from ground truth registries, frequently from websites such as Musicbrainz.org. Tools are employed to extract numerical features describing certain characteristics of the audio files. In the case of the experimental set-up used for the case study, we assume a more complex set-up where an external web service is used to extract such features. This forms the basis for learning a machine learning model using the WEKA machine learning software, which is finally employed to predict genre labels for unknown music. Further, several scripts are used to convert data formats and other similar tasks. The process described above can be seen as prototypical from a range of eScience processes, consisting both of external as well as locally available (intermediate) data, external web services as well as locally installed software used in the processing of the workflow, with several dependencies between the various components.

This scientific experiment has so far been executed by plugging together a number of Java programs, writing their data into intermediate files, and scripts implemented in the Linux shell to provide serial and parallel execution. This set-up does not provide a high degree of resilience to technological changes: the fact that both source data as well as ground truth data are provided externally does not allow the repetition of any experiment with comparable results. Dependencies on software and libraries installed in the experiment platform, that will usually change with frequent system updates further limit repeatability and re-executability. This is further threatened by the fact that the logic of extracting the numeric descriptors is encapsulated in an external service that may update its functional description at any point in time, potentially without providing any information on a service version update.

The scientific process as it is implemented and executed at the moment is exposed to a number of threats. For once, the process is not very well documented, e.g. the exact input and output parameters of each step are not defined, as well as the sequence of execution of the scripts. It is also dependant on the shell of a specific operating system. Even if e.g. the operating system and version is the same, local configuration of the default shell can vary for example on the Linux system, and thus scripts might be not be executable. Monitoring of the process execution is difficult, as there is no direct support available from the shell to capture input and output parameters. Finally, shell scripts might not be

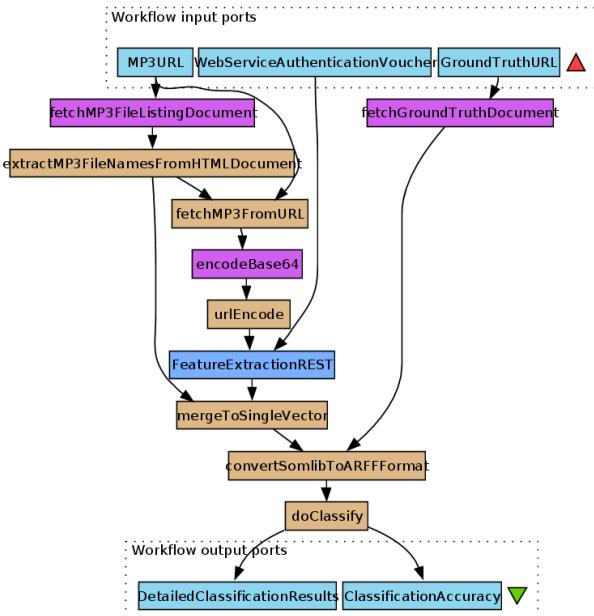


Figure 1: Scientific workflow modelled in the Taverna Workflow engine

persistently stored, but just typed and executed on the shell, and lost upon ending that shell session.

Even though individual of these aspects might be resolved with different means, migration of the process to a workflow management system seems to be a holistic approach towards the digital preservation process.

3.1 Implementation in Taverna

The implementation in Taverna required the migration of scripts and commands, that would have been executed with the shell of the operating system, to scripts in the Taverna-supported language (*beanshell*). These scripts were mainly used for performing format migrations for converting output data into the desired input format for the following tasks. The first step of the workflow could be solved by using services shipped with Taverna. We queried a directory list of a Web server containing the music files to be classified. After the server returned the HTML document containing the URLs of the files, a beanshell script was required to parse the actual locations of the files and for preparing a list. Consequently, the files had to be fetched from the collected URLs. This was achieved by modifying a provided Taverna service slightly to adapt it from images to MP3 files.

The next step (Base64 encoding) could be accomplished with an unmodified Taverna service, but had to be followed by a custom encoding beanshell for ensuring URL safety of the encoded files. The now correctly encoded files are passed to a REST-Web service in the following step. Taverna provides a ready to use REST invocation service, that has to be fed with the according file and an authentication voucher. After the service processed the files, a custom beanshell script was used for merging the single feature vectors to combined file.

The resulting file was then converted to the so called ARFF format, by using a beanshell script which invokes a third party library. This Java library had to be provided to the Taverna classpath in advance of the execution, and the usage of the library has to be explicitly specified in the beanshell service using it. After this has been achieved, the API of this library can be addressed via beanshell as if it were regular Java. The final step was again solved by using a beanshell script and an external library, which performs the actual classification.

The implementation in Taverna is fairly straight forward, as it allows to use the power of Java by a simplified scripting language. The library of existing local and remote services is extensive, and these services can easily be adapted to meet required specifications. Another advantage of Taverna is that the design, adaptation and execution of scientific experiments is integrated completely into the workbench, which reduces the installation and administration effort.

3.2 Implementation in Kepler

Kepler also provides scripting capabilities for Python, specifically by Jython⁵, an implementation which runs inside the Java Virtual Machine (JVM) and thus does not require additional software packages to be used. Nevertheless, as the third party libraries used in the process are written in Java, we were hinted at implementing the required actors in Java as well. This however is a serious overhead compared to being able to use the third-party library directly from beanshell as in Taverna. To implement the custom actor, one needs to set up the Kepler development environment. The build process is documented well, but requires several steps until the actual development of actors can be achieved.

The workflow implemented in Kepler is depicted in figure 2. A dynamic dataflow (DDF) director is used as there are several loops in the workflow. The first actor activated in the workflow is the Web Service. As invoking the service requires several steps, we encapsulated the internal logic into a so called composite actor. A composite actor itself contains a sub-part of the workflow and is used for reducing the complexity of the overall workflow, by hiding certain parts from the overview.

Although Kepler ships with a large amount of ready to use actors, it was necessary to implement several custom actors for e.g. Base64 and URL encoding on our own. The capabilities of wrapping existing actors, as it is enabled in Taverna, without modifying their source code is limited. Also the implementation of standard dataflow controls such as loops and conditionals requires many small intermediate steps, which render the overall process hard to read, interpret and understand.

3.3 Implementation in Activiti

After setting up the development environment in the Eclipse IDE, the implementation of the workflow in Activiti is rather straightforward, as task definition in the BPMN diagram and implementation of these classes are conveniently integrated. Even though Activiti does not provide a library

⁵www.jython.org/

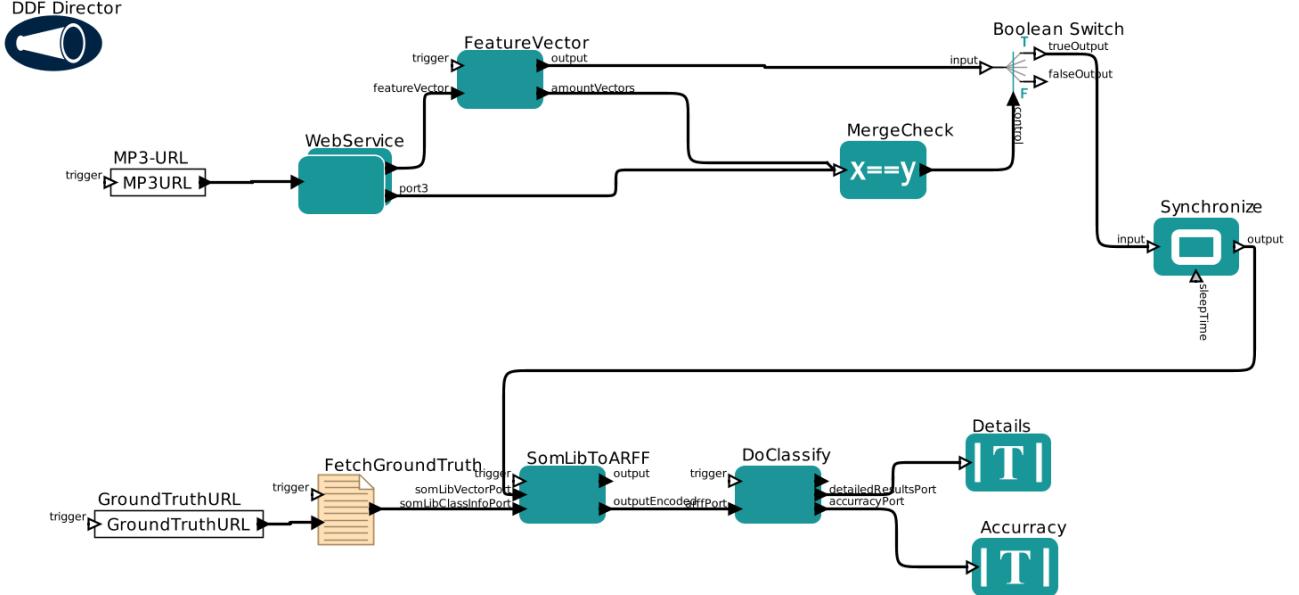


Figure 2: The Music Process Workflow designed in Kepler

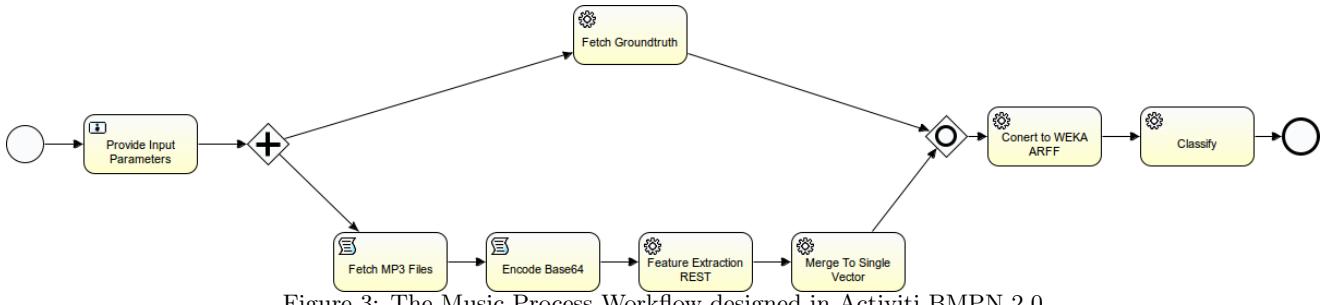


Figure 3: The Music Process Workflow designed in Activiti BMPN 2.0

of commonly repeating task implementations, the straightforward usage of Java as task implementation language allows to draw on the rich set of third-party libraries for compact implementations. Some steps, such as the fetching of files and encoding thereof, have been implemented with in Javascript. Fetching the genre assignment ground truth, calling the feature extraction REST service, converting the data format and performing the classification were solved as Java task implementations.

4. VALIDATION AND VERIFICATION OF PROCESS EXECUTION

Preserving workflows entails the requirement of validating their intermediate results and the overall output of a process execution. Preservation is only then considered a success, if all identified significant properties are equal before and after the preservation. The challenge of keeping workflows accessible is caused by the dynamic nature of processes. External components such as Web services and third party libraries are beyond the influence of workflow designers and users. These components might change at any point in time without prior announcement. Hence, they are critical threats to long term accessibility of workflows. In order to detect

these changes, it is necessary to monitor the workflow and the intermediate results they produce. This is a crucial requirement, as otherwise the reproducibility of workflows is at risk.

Measuring changes in significant properties is a difficult task. The authors of [3] propose a framework for evaluating whether two versions of a digital object are identical. The framework consists of several steps, that allow to identify significant properties of digital objects and examine their equivalence after an emulation process. These steps include the description of the original environment and the identification of external resources, that are beyond the influence of a system and influence the object to be preserved. The authors stress that there are different levels, at which objects can be compared with each other. This is also true for workflows. After a workflow has been preserved, the new environment has to be tested if it behaves the same way as the original. Thus test data needs to be used in order to extract and compare the significant properties of the workflow, i.e. if the reaction is identical to the data in both, the original and the preserved environment. The focus of [3] is on emulation as a preservation strategy. The underlying concepts can nevertheless be applied for other preservation strategies as well,

that is for instance migrating a workflow, and specifically its components, between different workflow engines.

When preserving processes, the data flow and the causal relationships between involved services can be seen as significant properties. Descriptions of this information is therefore required in order to compare intermediate results with each other. Most workflow management systems use the concept of provenance data to answer questions about execution and design details. Hence provenance data can be used for direct comparison of workflows across workflow engine boundaries.

4.1 Provenance Data

Provenance data describes the lineage of data and provides evidence about execution details, involved services and their intermediate results. A taxonomy of provenance techniques used in various WFMS was introduced in [12]. This taxonomy allows to categorize different systems based on the purpose of recorded provenance, their focus, representation, storage and dissemination. A further distinction between provenance systems can be achieved by the locus of data processing control as identified in [1]. The authors distinguish between command line based data processing, script and program based data processing, query based processing, service based processing and workflow management systems based processing. Depending on the type of data processing, different provenance data can be collected. Provenance data captured during process execution is thus an important aspect that must be captured as process context.

The recorded provenance data can be utilised to verify whether the new version of the process still renders the same results. To this end, as the evaluation framework suggests, one can automatically reapply the inputs and verify the recorded outputs, similar to what would be performed in automated software testing.

The provenance data can further be used for implementing a watch service for software and external service dependencies, e.g. by periodically executing the process with all historic recordings of previous executions, either as a complete process, or for each process step individually.

4.1.1 Provenance Capturing in Taverna

Taverna is capable of capturing the data exchanged between the process steps as provenance data, and stores it in a relational database (Apache Derby). Taverna records all invocations of the workflow and its individual steps, along with the data exchanged and timing information. The data can be exported in the Taverna-specific format Janus [8]; the also available Open Provenance Model format [10] contains only information of the invoked process steps, but not the actual data, and no information about execution time.

An example of the provenance data recorded for the two process outputs, the percentage of correctly classified instances, and the detailed classification results, are given in Listings 1 and 2 (note that some unique identifiers, such as URLs as namespaces, and identifiers for the workflow and specific data elements, have been abbreviated for space reasons).

Listing 1: Example provenance data of Taverna for the process output *ClassificationAccuracy* (cf. Figure 1). The first RDF Description element defines the output port *ClassificationAccuracy*, the second element contains the actual value of “80.0”.

```
<rdf:Description rdf:about="{nsTaverna}/2010/workflow/{idWF}/processor/
    MusicClassificationExperiment/out/
    ClassificationAccuracy">
<janus:has_value_binding rdf:resource="{nsTaverna}/2011/
    data/{idDataGrp}/ref/{idDataPort0}"/>
<rdfs:comment rdf:datatype="{nsW3}/2001/XMLSchema#string">
    "80.0"
</rdfs:comment>
<janus:is_processor_input rdf:datatype="{nsW3}/2001/
    XMLSchema#boolean">
    false
</janus:is_processor_input>
<janus:has_port_order rdf:datatype="{nsW3}/2001/
    XMLSchema#long">
    0
</janus:has_port_order>
<rdf:type rdf:resource="http://purl.org/net/taverna/
    janus#port"/>
</rdf:Description>

<rdf:Description rdf:about="{nsTaverna}/2011/data/{idDataGrp}/ref/{idDataPort0}"/>
<rdfs:comment rdf:datatype="{nsW3}/2001/XMLSchema#string">
    "80.0"
</rdfs:comment>
<janus:has_port_value_order rdf:datatype="{nsW3}/2001/
    XMLSchema#long">
    1
</janus:has_port_value_order>
<janus:has_iteration rdf:datatype="{nsW3}/2001/XMLSchema#string">
    []
</janus:has_iteration>
<rdf:type rdf:resource="http://purl.org/net/taverna/
    janus#port_value"/>
</rdf:Description>
```

Listing 2: Example provenance data of Taverna for the process output *DetailedClassificationResults* (cf. Figure 1). The first RDF Description element defines the output port *DetailedClassificationResults*, the second element contains the actual value, one entry for each file tested, with the actual class, the predicted class, and the confidence of the classifier in the prediction.

```
<rdf:Description rdf:about="{nsTaverna}/2010/workflow/{idWF}/processor/
    MusicClassificationExperiment/out/
    DetailedClassificationResults">
<janus:has_value_binding rdf:resource="{nsTaverna}/2011/
    data/{idDataGrp}/ref/{idDataPort1}"/>
<rdfs:comment rdf:datatype="{nsW3}/2001/XMLSchema#string">
    DetailedClassificationResults
</rdfs:comment>
<janus:is_processor_input rdf:datatype="{nsW3}/2001/
    XMLSchema#boolean">
    false
</janus:is_processor_input>
<janus:has_port_order rdf:datatype="{nsW3}/2001/
    XMLSchema#long">
    0
</janus:has_port_order>
<rdf:type rdf:resource="http://purl.org/net/taverna/
    janus#port"/>
</rdf:Description>

<rdf:Description rdf:about="{nsTaverna}/2011/data/{idDataGrp}/ref/{idDataPort1}"/>
<rdfs:comment rdf:datatype="{nsW3}/2001/XMLSchema#string">
    1 2:Hip-Hop 2:Hip-Hop 0.667 (3.359461)
    2 2:Hip-Hop 2:Hip-Hop 0.667 (3.294687)
    3 1:Classica 1:Classica 0.667 (2.032687)
    4 3:Jazz 3:Jazz 0.667 (2.536849)
    5 1:Classica 1:Classica 0.667 (1.31727)
    6 1:Classica 3:Jazz + 0.667 (3.46771)
    7 3:Jazz 1:Classica + 0.333 (2.159764)
    8 2:Hip-Hop 2:Hip-Hop 0.667 (3.127645)
    9 3:Jazz 3:Jazz 0.667 (3.010563)
    10 2:Hip-Hop 2:Hip-Hop 0.667 (4.631316)
</rdfs:comment>
```

Each listing contains two RDF *Description* elements, where the first one defines the output port, and contains as a sub-

element the identifier of the element containing the actual value, which is the second *Description* element in both listings. With the identifiers used in the *rdf:about* attributes, it is possible to uniquely identify the process step (and iteration, if the step is looped over) the data originates from.

4.1.2 Provenance Capturing in Kepler

The Kepler SWMS provides a dedicated module for recording provenance information[4]. When this software component is loaded, a specialized actor called *Provenance Recorder* is available. This actor is used for monitoring the process execution and storing metadata about the workflow persistently. The provenance module stores provenance data by default in the relational database HyperSQL (HSQLDB⁶). It is integrated directly into the provenance module and can be queried by using the Database Manager provided by HSQLDB. Kepler stores detailed metadata about every execution of a workflow. This covers actors, ports, parameters, relations and additional information about the workflow, such as user names and context information. The Kepler provenance system also stores the data used during the execution, which allows the detection of changes within the results.

All information stored within HSQLDB can be queried by using standard SQL, and from a Java program via an API. The OPM export feature completes the provenance data management of Kepler; in contrast to Taverna the exported OPM XML file contains time stamps and allows to derive the execution sequence easily.

Listing 3: A Kepler OPM XML snippet

```
<wasGeneratedBy>
  <effect id="_a2"/>
  <role value="output"/>
  <cause id="-p0"/>
  <time>
    <noLaterThan>16:26:17.333+02:00</noLaterThan>
    <noEarlierThan>16:26:17.333+02:00</noEarlierThan>
    <clockId>-c1</clockId>
  </time>
</wasGeneratedBy>
```

Listing 3 depicts an example of an exported OPM file. It contains references to the actor that generated the output (-p0) and refers to the the output of this event (_a2), using auto-generated identifiers to refer to these elements.

4.1.3 Provenance Capturing in Activiti

Activiti refers to provenance data as (process execution) history, and allows to configure recording on several levels of detail. Similar to the other systems, the data is stored in a relational database (H2⁷), which can be queried to retrieve information about the process and task invocations. As there is no explicit input and output of process steps (ports in Taverna and Kepler), rather the global state of data in the process execution is stored, than specific parameters for a specific task invocation. Activiti also does not provide an export into e.g. the OPM format.

⁶www.hsqldb.org/

⁷<http://www.h2database.com>

5. COMPARISON OF WORKFLOW MANAGEMENT SYSTEMS

We identified a number of criteria important for the migration and execution of processes workflow management. A summary of these criteria is provided in Table 1.

Regarding setup of the design and execution platform, Kepler and Taverna provide a straightforward installation routine, while Activiti requires a bit more work with preparing the Eclipse IDE and plugins.

All systems evaluated in this paper allow to implement the process with the use of the Java programming language, even though the complexity of doing so differs greatly; both Kepler and Taverna require the programmers to develop the modules outside the workflow management system and then to register their services and actors, respectively, with the engine. Implementing tasks in Activiti benefits from the initial setup of the Eclipse environment.

The systems differ greatly when it comes to the support of scripting languages for fast and simple tasks. Here, Activiti provides the widest range of languages, and is in theory not limited, as long as the language conforms with the Java Scripting Platform. If there are a lot of legacy scripts that would need to be preserved, Activiti would thus seem to be a prime choice. It seems vital that other systems would allow such a wide range of implementations as well. Still, this will also raise the complexity of preserving the actual process as components in many different languages may need to be preserved, together with their operational infrastructure (compiler, interpreter, runtime environments, etc.). Kepler provides Python, and Taverna Beanshell scripting capabilities. The latter further provides a large library of services that can be used to quickly perform common tasks, and allows to easily alter these template implementations.

All systems allow to record provenance data during the process execution, which enables for validation. Kepler and Taverna provide various types of exports of this data, in the Open Provenance Model (OPM) and custom formats, and are more detailed on the single processing steps, as input and output ports of each process step are clearly defined. This seems to be an important aspect for detailed validation and watch activities. Activiti could be easily augmented by an export into the OPM, and input and output parameters for a processing step could, for a specific process execution, be deduced from the change in global variables.

6. CONCLUSIONS

The preservation of complete (business) processes is starting to be understood as a new challenge in digital preservation research. Scientific processes need to be preserved to allow later verification of results, and business process preservation can play an important role when it comes to liability or patent infringement litigations.

However, process preservation is inherently difficult – today's processes are executed inside complex software ecosystems, and composed of a myriad of services. Capturing this software setup and its configuration is only one step – without being able to validate the process execution, we cannot

Table 1: Features of Workflow Management Systems

Engine	Implementation	Script Language Support	Designer Support	Execution Engine	Provenance Capturing	Provenance Export
Taverna	Java	Beanshell (Java)	Standalone	Integrated with designer	Database (Apache Derby)	OPM & Janus
Kepler	Java	Python	Standalone	Integrated with designer	Database (HSQLDB)	OPM
Activiti	Java	JavaScript, Python, Ruby, Groovy, ...	Via Eclipse IDE	Web application or Java program	Database (H2 DB)	-

guarantee that the preserved process is still the same when re-executed at a later time.

These two concerns are a bit relaxed when defining and executing the process in a dedicated workflow engine, which provides a layer of abstraction to the original software setup. It also allows to closely monitor, and thus evaluate, the process execution. In this paper, we therefore described a number of popular workflow management systems, and described how they can be used to migrate a scientific experiment process. Efforts for migrating a workflow to a workflow management system might be significant; therefore, flexibility in the implementation is a prime aspect.

With the migration of a process to a workflow management engine, we can mitigate a few concerns that can hamper the preservation of this process. First, the migration to workflow engines has the benefit of requiring a clear and formal definition of the processes, which might not be present before. Thus, we obtain documentation and detailed descriptions on the process. Further, we can evaluate and monitor the execution of the processes closely, which enables verification that a process is still executed unchanged. Finally, the migration to a workflow management system in general is a step of abstraction from a specific software setup. The requirements and interfaces to operating systems or and system libraries are greatly reduced, and dependencies on third-party libraries are generally explicitly defined.

The migration does not prevent external elements such as the webservice employed in our case study from becoming obsolete. Thus, contracts and service level agreements have to be agreed on with the providers of these services to maintain and migrate their services if needed. Then, using previously recorded provenance data, we can verify whether these services still behave as before.

7. ACKNOWLEDGMENTS

Part of this work was supported by the projects APARSEN and TIMBUS, partially funded by the EU under the FP7 contracts 269977 and 269940.

8. REFERENCES

- [1] R. Bose and J. Frew. Lineage retrieval for scientific data processing: a survey. *ACM Comput. Surv.*, 37(1):1–28, Mar. 2005.
- [2] J. Freire, D. Koop, E. Santos, and C. T. Silva. Provenance for computational tasks: A survey. *Computing in Science and Engg.*, 10(3):11–21, May 2008.
- [3] M. Guttenbrunner and A. Rauber. A Measurement Framework for Evaluating Emulators for Digital Preservation. *ACM Transactions on Information Systems (TOIS)*, 30(2), 2012.
- [4] The Kepler Project. *Getting Started with Kepler Provenance 2.3*, August 2011.
- [5] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao. Scientific workflow management and the kepler system. *Concurrency and Computation: Practice and Experience*, 18(10):1039–1065, 2006.
- [6] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao. Scientific workflow management and the kepler system: Research articles. *Concurr. Comput. : Pract. Exper.*, 18(10):1039–1065, Aug. 2006.
- [7] Music Information Retrieval Evaluation eXchange (MIREX). Website. <http://www.music-ir.org/mirex>.
- [8] P. Missier, S. S. Sahoo, J. Zhao, C. A. Goble, and A. P. Sheth. Janus: From workflows to semantic provenance and linked open data. In *Proceedings of the International Provenance and Annotation Workshop (IPA 2010)*, pages 129–141, Troy, New York, USA, June 15–16 2010.
- [9] P. Missier, S. Soiland-Reyes, S. Owen, W. Tan, A. Nenadic, I. Dunlop, A. Williams, T. Oinn, and C. Goble. Taverna, reloaded. In M. Gertz, T. Hey, and B. Ludaescher, editors, *SSDBM 2010*, Heidelberg, Germany, June 2010.
- [10] L. Moreau, J. Freire, J. Futrelle, R. E. McGrath, J. Myers, and P. Paulson. *Provenance and Annotation of Data and Processes*, chapter The Open Provenance Model: An Overview, pages 323–326. Springer-Verlag, Berlin, Heidelberg, 2008.
- [11] C. Silva, J. Freire, and S. Callahan. Provenance for visualizations: Reproducibility and beyond. *Computing in Science Engineering*, 9(5):82 –89, sept.-oct. 2007.
- [12] Y. L. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. *SIGMOD Rec.*, 34(3):31–36, Sept. 2005.
- [13] A. van der Aalst, A. Hofstede, and M. Weske. Business process management: A survey. In M. Weske, editor, *Business Process Management*, volume 2678 of *Lecture Notes in Computer Science*, pages 1019–1019. Springer Berlin / Heidelberg, 2003. 10.1007/3-540-44895-0_1.

Digital Preservation Of Business Processes with TIMBUS Architecture

Mykola Galushka SAP (UK) Ltd The Concourse, Queen's Road, Queen's Island, Titanic Quarter, Belfast, BT3 9DT mykola.galushka@sap.com	Philip Taylor SAP (UK) Ltd The Concourse, Queen's Road, Queen's Island, Titanic Quarter, Belfast, BT3 9DT philip.taylor@sap.com	Wasif Gilani SAP (UK) Ltd The Concourse, Queen's Road, Queen's Island, Titanic Quarter, Belfast, BT3 9DT wasif.gilani@sap.com
John Thomson CMS Portugal john.thomson@caixamagica.pt	Stephan Strodl Secure Business Austria Vienna, Austria sstrodl@sba-research.org	Martin Alexander Neumann KIT, Kaiserstrasse 12, 76131 Karlsruhe, Germany mneumann@teco.edu

ABSTRACT

The majority of existing digital preservation solutions are focusing on the long-term storage of digital content such as documents, images, video, audio files and other domain specific data. Preservation of an Information Technology infrastructure for supporting business processes is a much more challenging task. It requires the preservation of software and hardware stacks as well as relevant contexts, which together, provide an execution layer for running business processes. The proposed TIMBUS architecture addresses limitations of existing digital preservation solutions and provides a complete framework for preserving business processes implemented upon a service oriented architecture.

Keywords

digital preservation, business process

1. INTRODUCTION

Providing business continuity is an important task facing many multi-national companies. Failure to provide business continuity often leads to substantial financial losses and may cause total or partial loss of business. The most common approach to address potential adverse circumstances or events is to provide effective Business Continuity Management (BCM). The main objective of BCM is to access different risk factors and establish contingency plans for minimising impact of potential hazardous events on existing business processes (BPs).

Many BPs are heavily dependent on sophisticated informa-

tion technology (IT) infrastructure, which is supporting different business transactions and providing critical information for making important business decisions. Constant evolution of the technology landscape brings new challenges for any business organisation to support their IT infrastructure, where new hardware and software solutions are developed and adopted on a regular basis. These new versions of hardware and software components often have backward compatibility limitations. In some contexts digital preservation (DP) is one of the most effective solution for dealing with evolving digital infrastructures.

The majority of existing DP solutions focus on the long-term storage of digital content such as documents, images, video and audio files. Preservation of an IT infrastructure for supporting BPs is a much more challenging task. It requires the preservation of software and hardware stacks as well as relevant contexts, which together, provide an execution layer for running BPs.

DP architecture developed within the TIMBUS¹[12] project provides a unique set of solutions going beyond the scope of existing DP approaches. It covers all aspects of traditional DP system such as preserving a digital content but also addresses enterprise risk analysis and business continuity planning. It covers a wider scope of DP processes, which includes Intelligent Enterprise Risk Management (IERM) for automatic identification and prioritisation of risks within an enterprise and ability to minimize those risks by taking a specific set of actions including DP.

The TIMBUS system identifies a set of interdependent BPs from the enterprise logs, automatically detects and captures relevant context meta data, packages the collected information and provides facilities for long term preservation, monitoring and maintenance. The TIMBUS system enables the redeployment and re-execution of the partial or complete BP at a future time.

¹<http://timbusproject.net/>

A high level overview of relevant research projects covering various aspects of DP, commercial and open source solutions for performing DP is presented in the next section.

1.1 Related Work

The EC has long history of supporting DP research SCAPE [20], WF4EVER [2], PROTAG [19], APARSEN [3] and LIWA [16]. Many of the completed and currently running projects have made important contributions and are relevant to TIMBUS. SCAPE is developing a services based framework for preservation of large-scale, heterogeneous collections of complex digital objects by using semi-automated work-flows based on open source platforms. WF-4-EVER is developing a software architecture for the preservation of scientific work-flows for conducting complex research experiments in combination with relevant contexts. A smart multi-agent architecture for performing the long-term preservation of digital objects is developed in PROTAG project. It can be integrated with existing and new preservation systems to support various aspects of a DP work-flow. A sustainable digital information infrastructure for supporting permanent access to digitally encoded information is developing based on the APARSEN framework. LIWA has developed an architecture for Web content preservation, which supports capturing content from a wide variety of sources and performs the long term interpretation of constantly evolving data archive by filtering out irrelevant information.

The following projects have some degree of overlap with DP, where the main focus lies in providing framework for the long term access to particular information resources. In the scientific domain, projects like GENESI-DR[15] and PREPARINGDARIAH[18] provide infrastructure for establishing an open digital repository for world-wide researchers to seamlessly access and share data, information and knowledge originating within different areas of science. Long term preservation in libraries and museums is investigated by AR-COMEM[4], PATHS[17], AXES[1], PAPYRUS[8] and DECIPHER[11] projects. ARCOMEM is focusing on transforming digital archives into memories structures, that can be utilised by specific community of experts, where PATHS is implementing an approach for interpreting heritage material and providing clear navigation tools for wide range of users. AXES is developing a set of tools for providing intelligent interactions with various types of digital content. An innovative ideas of understanding user queries in the context of different specific disciplines for improving underlying search techniques are developed in PAPYRUS project. Using of semantic web technologies for analysing digital heritage are investigating in DECIPHER project. ENSURE[13] provides the long-term usability of data produced or controlled by commercial organisations.

Processing of digitally preserved data are explored in the following projects SOAP[21], CULTURA[7], 5-COFM[14], ETHIO-SPARE[10], ARTSENSE[5], CHESS[9] and BLOG-FOR-EVER[6]. Innovative approaches to filtering, restructuring and facilitating experience of interaction with scientific publication are developed within the scope of SOAP and CULTURA projects. Information discovery aspects in digital archives relevant to the cultural heritage domain are investigated in 5-COFM, ETHIO-SPARE and ARTSENSE projects. User engaging techniques for providing better and

more efficient access to historical and cultural information as well as modern blogs are developed in CHESS and BLOG-FOR-EVER projects.

There are a number of commercial and open source products available, which address different aspects of DP. SDB² developed by world leading company in DP technologies provides services for storing and preserving critical digital information in reliable manner. DIAS³, developed by IBM, addresses various aspects of long-term usability of digital information over the past decade. Rosetta developed by ExLibris⁴ provides a highly scalable, secure, and easily managed DP system for preserving knowledge, libraries and other memory institution data around the world.

Alongside the commercial applications open source projects are available: Fedora-Commons, Greenstone, LOCKSS, Archimatica, DPSP, IRODS, DAITSS. CDS Invenio & CERN⁵ Document Server supports preservation of articles, books, journals, photos, videos etc. and used by large number of scientific institutions worldwide. DSpace⁶ developed by the Massachusetts Institute of Technology Libraries and Hewlett-Packard supports building of open digital repositories for publishing content. Eprints⁷ is a set of open-source software applications for building open access services to publishing and multimedia content, which support a number of features such as meta-data extraction, access control, flexible work-flows etc.

The large verity of research, commercial and open source, in area of DP shows that the problem of DP is well-understood for data-centric information scenarios. On the other hand, scenarios where important digital information has to be preserved together with the execution contexts have been less explored. Preservation is often considered as a set of activities carried out in isolation within a single domain, without taking into account the dependencies on third-party services, information and capabilities that will be necessary to validate digital information in the future. Many existing DP solutions focus on more simple data objects which are static in nature. The unique aspect of TIMBUS is that it attempts to advance state of the art by exploring how more complex digital objects can be preserved, such as BPs with the entire execution environment. TIMBUS provides the infrastructure, which supports the user in identifying what BPs to keep, why and for how long they need to be kept.

Many modern BPs are exposed to the outside world via service oriented architectures (SOA). SOA is one of the most popular approaches, used by modern companies, for facilitating their business activities over the Internet. It provides a fast, reliable and convenient way to reach a large volume of customers world-wide. The long-term preservation of SOA based solutions is a critical necessity faced by many companies to ensure some level of business continuity. The focus on DP of BPs, where SOA is used as a framework for deliv-

²<http://www.digital-preservation.com/>

³<http://www-935.ibm.com/services/nl/dias/>

⁴<http://www.exlibrisgroup.com/>

⁵<http://invenio-software.org/>

⁶<http://www.dspace.org/>

⁷<http://www.eprints.org/>

ering services, is the main aim the TIMBUS project. Since understanding of SOA is critical for the successful design of the TIMBUS DP architecture, a high level overview is presented below.

1.2 Service Oriented Architecture

The term "service" defines a system of organised resources used for supplying specific needs to particular individuals or organisations. Services address relevant concepts from different domains [22] such as economy, business, science, etc. A typical service-based model in business [24] combines the following three service layers: *Business Service*, *E-Service* and *Web-Service*.

- *Business Service* represents the non-material equivalent of goods. They are defined as a set of activities supplied by service providers to service customers in order to deliver a specific set of values. Traditionally the majority of these services are discovered and invoked manually, while their realisation may performed by manual, semi-automated or automated fashion.
- *E-Service* is provided and executed by electronic systems in an automated fashion. IT provides an infrastructure for developing concepts such as e-service or e-commerce (electronic-commerce). Such services are executed via transactions conducted over the Internet or an internal computer network. These on-line transactions include buying and selling goods, where business is done via Electronic Data Interchange (EDI). EDI is performed by a collection of software components communicated via standardised network protocols.
- *Web Service* is the e-service consumed via Web-based protocols or Web-based programs. Separation of logical and technical layers gives a possibility of using alternative technologies for the e-service implementation. The following three types of Web-service architectures can be identified: *RPC Web Service* [23, 28, 30], *SOA Web Service* [26] and *RESTful Web Service* [29].

These services can be combined into two main business models used by service providers: *Software as a Service* (SaaS) [25] and *Internet of Services* (IoS) [24].

The SaaS Model is characterised by the following factors: provides a quicker-to-deploy strategy, quicker return of investment, frequent and automatic updates, independence from other IT components and improved usability; provides a low-risk alternative to traditionally licensed software; makes business units more focused on business transactions by eliminating dependence on supporting an IT infrastructure; facilitates a collaborative development of complex business models.

The initial focus of SaaS on the middle size companies has been expanded to the enterprise level, which changed the overall software applications market. It requires software vendors to carefully adjust their offer for meeting the constantly rising customer demand on SaaS solutions.

The IoS Model extends the concept of SaaS by providing mechanisms for discovering and invoking new services. It includes a variety of components such as standards, tools and applications for supporting business transactions. These components bring together service providers and consumers in the service-market place, where they can be more efficiently engaged in the verity of business activities.

IoS is also focusing on the creation of business networks, where elements of these networks could support SaaS models. IoS therefore provides infrastructures such as marketplace, brokerage, integration, interoperability, aggregation etc. for multiple services based on the SaaS model.

Rapid adoption of SaaS and IoS models indicates that many BPs are built on service-oriented architectures. Numerous services can be provided by different providers and operated from different geographical locations. A composition of outputs provided by each individual service is combined into particular business value, which can be utilised further by service consumers. Despite the clear advantages of SaaS and IoS, there is a danger of disappearing services and service providers (due to various reasons) by leaving some BPs partially incomplete. Considering that business continuity is not only a company desirable requirement, but also frequently a legal obligation, DP of BPs becomes as important factor of the modern business strategy. Service providers will ultimately be responsible for incorporating TIMBUS-like preservation solutions into future offerings, to support the long-term sustainability of their business models. The TIMBUS architecture presented in the next section provides the complete solution for preservation of complex BPs.

2. ARCHITECTURE

A high-level view of the TIMBUS DP architecture is shown in Figure 1. It consists of five modules: *DP Agent*, *DP Acquisition*, *Intelligent Enterprise Risk Management (iERM)*, *Legality Life-cycle Management (LLM)* and *DP Engine*.

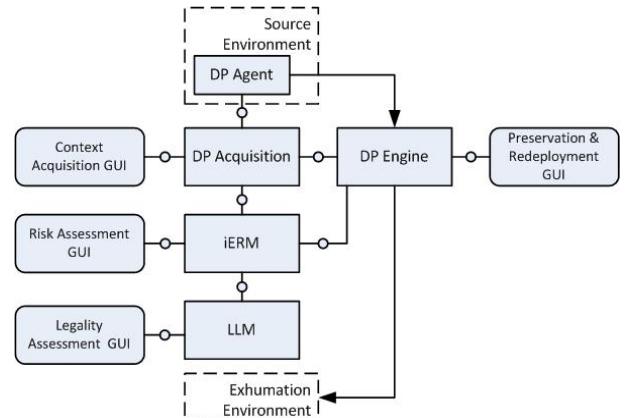


Figure 1: The high-level view of the TIMBUS DP architecture.

The *DP Agent* Module is running within the *Source environment* and capturing resources required for performing DP. The collected resources are utilised by the *DP Acquisition* Module, which extracts relevant contexts and combines them into a *Context Model Instance (CMI)*. This model is

annotated in *IERM* and *LLM* Modules with different risk factors and assessed according to the specified annotations. A generated risk assessment report is used by the *DP Engine* Module for selecting the most suitable preservation strategy. The selected strategy is translated into a preservation plan, which includes the complete set of instruction for execution process. During a redeployment phase the *DP Engine* Module identifies a difference between preserved and currently available environments. The identified difference is used for planning and execution by a redeployment process. All steps of preservation and redeployment processes can also be verified by a specific subset of components integrated into the *DP Engine* Module.

The DP system is integrated with *Source* and *Redeployment* environments. *Source* environment is a combination of all IT and non-IT related resources, which support the execution of BPs and need to be fully or partially preserved according to identified risk factors. The source environment consists of all infrastructure and software components required at run-time. Context information is also required for future usability, including but not limited to dependencies between business process components and the business process itself. *Redeployment* environment is a combination of IT and non-IT related resources forming an infrastructure which supports the execution of archived BPs, which can be fully or partially redeployed based on information stored within the *DP System Archive*.

Preservation and redeployment processes are controlled via four graphical user interfaces (GUI): *Contexts Acquisition*, *Risk Assessment*, *Legality Assessment* and *Preservation & Redeployment*. The *Contexts Acquisition* GUI controls processes of contexts mining and creation of CMI. The *Risk Assessment* and *Legality Assessment* GUIs control the annotation of CMI and risk impact assessment processes. *Preservation & Redeployment* defines the core set of tools for controlling planning, execution and validation processes.

2.1 DP Agent Module

The *DP Agent* Module (see Figure 2) is a combination of software components, which are running within the *Source* environment and capturing resources required for performing DP.

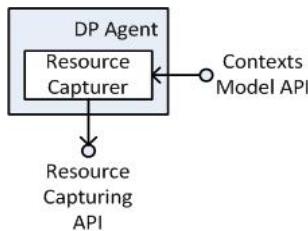


Figure 2: The architecture of the DP Agent Module.

The agent's key component *Resource Capturer* has a transparent, plug-in type architecture, which allows connecting to different resource capturing components for various IT systems. There are four main plug-in components included into the initial design: *Static Dependency Capturer*, *Dynamic Dependencies Capturer*, *Contexts Capturer* and *Event-logs Capturer*.

These plug-in components collect the key set of data required for carrying out further steps of DP. Obtained data are packaged before transferring to the DP Acquisition Module. They can be archived and/or encrypted to provide the required level of transferring efficiency and security. The created data package is transferred to the *DP Acquisition* Module via the secure peer-to-peer connection.

Control of *DP Agents* is performed via the secure communication channel, which is used for exchanging control messages. It allows control of the resource capturing process on multiple instances of the *DP Agents* from the single instance of the *DP Acquisition* Module. Such approach significantly simplifies a deployment process and minimises a cost of managing the large scale IT systems.

2.2 DP Data Acquisition

The *DP Acquisition* Module (see Figure 3) is a combination of software components, which are used for collecting and combining dependencies, contexts and event-logs from different *DP Agents* into the CMI. It consists of three software components: *Contexts Miner*, *Contexts Monitor* and *Model Weaver*.

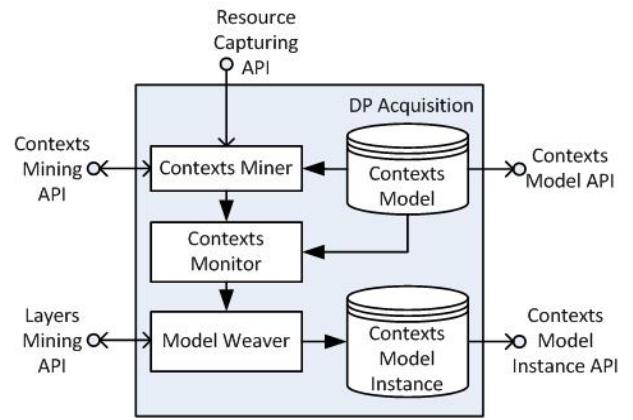


Figure 3: The architecture of the DP Acquisition Module.

The *Contexts Miner* collects data from different *DP Agents* and performs extraction of different types of information defined by the *Context Model* (CM). At the high-level of abstraction the mining process can be split into three parts: dependencies extraction, contexts extraction and BPs extraction. The dependency extractor identifies a set of software and hardware components operating within an IT landscape and establishes relations between them. The contexts extractor discovers relevant context information relevant to the BPs, which need to be taken into consideration during the DP process. The BPs extractor analyses enterprise event-logs and extracts distinct BPs and their interrelations.

The *Contexts Monitor* receives data discovered during the mining step and performs their evaluation. The main purpose of this evaluation is to establish whether or not discovered resources contain a critical amount of data required to initiate regeneration of CMI. For example, let's assume a monitoring database hosting customer data. If a sales team member introduces a new customer record, it is not going

to be considered as the critical event requires regeneration of CMI. However, if a member of the IT supporting team modifies the database design, then it may cause a significant impact on business logic regeneration of CMI for capturing introduced changes.

The *Model Weaver* combines discovered dependencies, relevant contexts and BPs into a single CMI. A few representational schemes for the CM were analysed within the TIMBUS Project. *Web Ontology Language*⁸ (OWL) was selected as the most suitable model for representing the unified view on collected resources and BPs. The OWL standard combines different consecutive approaches from the Semantic Web community. OWL provides a more expressive way to define relation mappings between resources and BPs discovered within the TIMBUS project than any other relevant model. Resources identified during the mining process are represented in OWL by entities. Entities can be further sub-divided into classes and instances. Class represents an abstraction, which combines instances with the common type. Each class aggregate is a common subset of properties shared between encapsulated instances. Decisions of subdividing instances are carried out during a model designing phase. Relations between entities are labelled by descriptive terms, which allow to form meaningful connections between two or more element and perform reasoning queries.

As a result of the data acquisition operations generated, CMI contains all necessary components for performing risk assessment carried out by the *IERM* Module, which is described in the next section.

2.3 Intelligent Enterprise Risk Management Module

The *IERM* Module (see Figure 4) is a combination of software components, which are used for assessing risks associated with BPs and dependent resources. This module generates a report describing risk levels and cost values associated with failure of particular subset of BPs. It consists of four software components: *Risk Model Builder*, *Risk Annotator*, *Risk Impact Assessment* and *Risk Monitor*.

The *Risk Model Builder* allows an expert to populate *Risk Model* with relevant risk factors to a specific subset of BPs. For instance, if a particular business is heavily dependent on consumption of natural gas, the relevant risk factors will include financial losses due to supply shortages, fluctuation of market prices due to political and economical situation, natural disaster etc. When relevant risk factors are defined, the *Risk Annotator* allows the user to assign them to BP instances defined CMI. This process is performed in a semi-automatic fashion, where IERM prompts the most suitable risk factors for particular BP instance and then an expert accepts or modifies the proposed suggestion. Annotated CMI components are formed into the *Unified Risk Model* (URM). URM is a sub-model of CMI, which is only focused on supporting simulation operations. The *Risk Impact Assessment* tool uses URM to assess the impact of different risk factors on BPs, business objectives and Key Performance Indicators (KPIs). The risk impact assessment is performed by constructing a Petri Net Model[27] and running simulation

⁸<http://www.w3.org/TR/owl-features/>

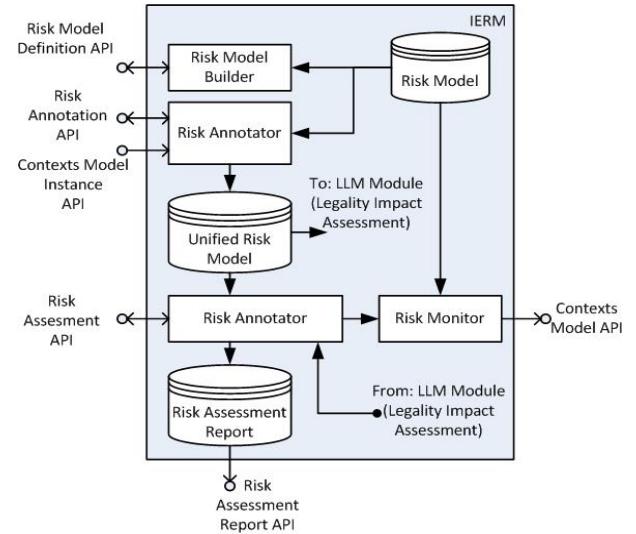


Figure 4: The architecture of the DP IERM Module.

processes. Result of these simulations are analysed to find weak points in examined BP models and compiled into the *Risk Assessment Report* (RAR).

The *Risk Monitor* constantly monitors the *DP System Archive* and CMI. It tries to detect any changes within the monitoring environment, which may lead to appearance of risk events. If the risk event is detected the monitor triggers the simulation process carried out by the *Risk Impact Assessment* and *Legality Impact Assessment* tools. The legality impact assessment is performed by the *LLM*, which is presented in the next section.

2.4 Legality Life-cycle Management Module

The *LLM* Module (see Figure 5) is a combination of software components, which are used for assessing impacts of legal issues on BPs. This module consists of two software components: *Legalities Annotator* and *Legality Impact Assessment*.

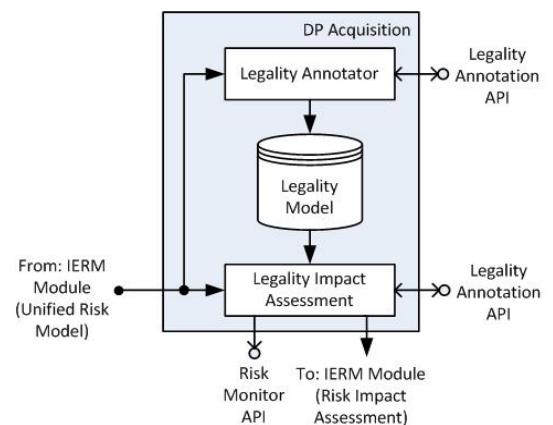


Figure 5: The architecture of the DP LLM Module.

The *Legalities Annotator* allows an expert to define legal and contractual issues relevant to a particular organisation or

project and check whether or not relevant resources and BPs from the URM are compliant with these issues. Considering the high complexity of this task the majority of annotation operations are supervised by an expert. Annotated URM elements are placed into the *Legalities Store*, which is created to support the legality impact modelling process.

When the *Legalities Store* is populated the *Legalities Impact Assessment* tool checks whether or not stored legality rules are enforced for the discovered set of BPs defined in URM. Results of this assessment are sent back to the *Risk Impact Assessment* tool in the IERM module, which takes the calculated legal risk into consideration and uses it in conjunction with other risk factors.

The RAR created by *IERM* and *LLM* contains important information about BPs and various level of risks and cost associated with their failure. This information is utilised by *DP Engine* described in the next section, which identifies the most suitable preservation and redeployment strategy and performs its execution.

2.5 DP Engine

The *DP Engine* Module is a combination of software components, which are used for generating preservation and redeployment plans by utilising the RAR and CMI. This module also provides mechanisms for verification and testing different stages of preservation and redeployment processes. The *DP Engine* consists of three distinct cycles: *Preservation*, *Redeployment* and *Verification & Feedback*.

2.5.1 Preservation Cycle

The preservations cycle (see Figure 6) includes elements for preparing the preservation plan and performing its execution. It consists of the following elements: *Preservation Alternatives Assessment*, *Preservation Execution Planner*, *Process Preservation Executor*, *Preservation Monitor* and *DP System Archive*.

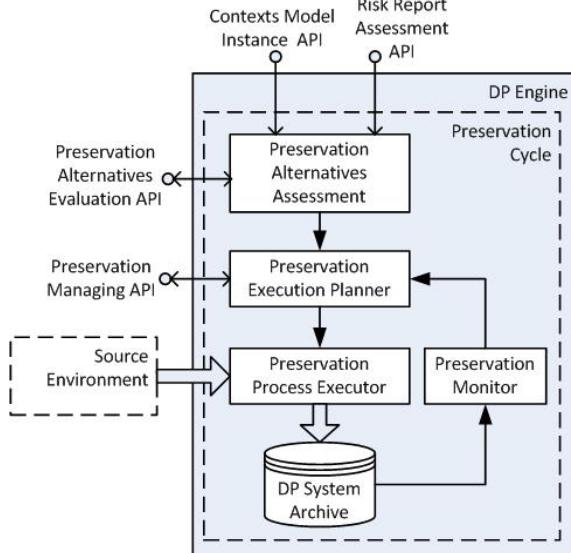


Figure 6: The architecture of the DP Engine Module for Preservation Cycle.

The *Preservation Alternatives Assessment* component analyses the generated RAR and CMI to identify the most suitable preservation strategy for most critical BPs. The assessment process takes into consideration risks associated with each BP and identifies the most effective preservation strategy. The strategy selection is performed in collaboration with an expert in an interactive mode, where the system provides different preservation alternatives and gives the expert priority to make the final decision. The selected strategy is analysed and converted into a list of instructions by the *Preservation Execution Planner*. These instructions can be executed in manual, semi-automatic or automatic fashion depending on complexity of the underlining processes. Execution of prepared instructions is carried out in the *Process Preservation Executor*. If an interaction is related to preservation of a specific IT entity, then the executor performs the automatic extraction of the requested components form the underlying IT landscape. It is followed by a set of transformation and packaging operations, required for long-term storing of selected IT entities in the *DP System Archive*. All processes are closely monitored by *Preservation Monitor*. It detects any deviations from the original script and notifies the *Preservation Execution Planner*, which initiates a re-planing phase. All decisions made during the preservation cycle are logged together with preservation steps in the specific log-containers within the *DP System Archive*.

2.5.2 Redeployment Cycle

The Redeployment cycle (see Figure 7) includes elements for preparing the redeployment plan and performing its execution. It consists of the following elements: *Redeployment Alternative Assessment*, *Redeployment Execution Planner*, *Process Redeployment Executor*, *Redeployment Monitor* and *DP System Archive*.

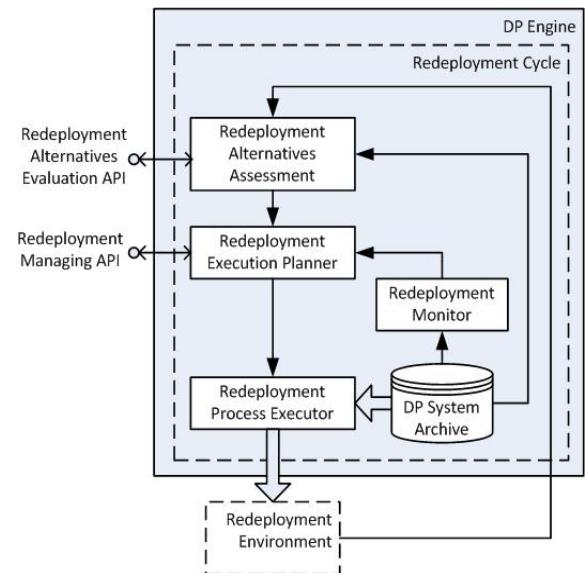


Figure 7: The architecture of the DP Engine Module for Redeployment Cycle.

The redeployment cycle performs the opposite operations to preservation. The *Redeployment Alternative Assessment* collects the information about IT landscapes, one which is

initially preserved and another which is currently running within the *Redeployment* environment. Base on the collected information it identifies missing components in the currently running environment and performs assessment and selection of the best redeployment strategy. The selected strategy is utilised by the *Redeployment Execution Planner*. It generates a list of actions which exhume the selected IT components to the redeployment environment. It is important to mention that not all steps in the redeployment plan are IT related. Some of them may required verification of particular legal issues, installing the specific hardware equipment etc. The prepared redeployment plan is passed to *Process Redeployment Executor* which works in a semi-automatic fashion. It automatically executes steps for redeploying IT entities from the *DP System Archive* to the *redeployment* environment. For all other steps, which cannot be completed without an external input, it provides an assisting interface, which guides execution of these steps. All redeployment processes are closely monitored by *Redeployment Monitor*. It detects any deviations from the original script and notifies the *Redeployment Execution Planner*, which initiates a re-planing phase. All decisions are made during the redeployment cycle are logged together with redeployment steps in the specific log-containers within the *DP System Archive*.

2.5.3 Validation and Feedback Cycle

The Validation and Feedback cycle (see Figure 8) includes elements for verification and testing the preservation redeployment processes. It consists of the following elements: *Preservation Log Gap Detector* and *Validation & Feedback*.

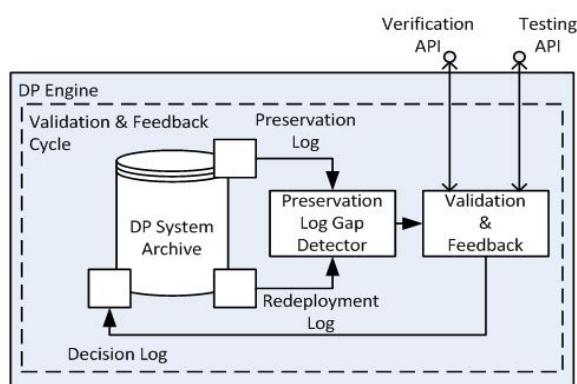


Figure 8: The architecture of the DP Engine Module for Validation & Feedback Cycle.

The *Preservation Log Gap Detector* compares the preservation and redeployment logs on potential inconsistencies. These inconsistencies may occur due to a variety of different reasons such as failures with hardware or software components during preservation or redeployment phases, unresolved legal issues, execution failures of manual steps etc. The collected information about potential issues is supplied to the *Validation & Feedback* component, which generates a set of instructions for resolving occurred issues. Considering the high complexity of this system, all operations in this phase are supervised by an expert. Another important function supported by this cycle is testing and rerunning of preserved BPs. It allows a continuous evaluation of preserved BPs and detects potential problems with the redeployment

process in advanced.

3. CONCLUSIONS

The proposed TIMBUS DP Architecture provides a complete framework for preserving BPs with all relevant dependencies and contexts. It has a generic, transparent and extendable design, which overcomes limitations in existing DP solutions. It is fully compliant with all requirements and use-case scenarios developing within the TIMBUS project. These use-case scenarios cover preservation of complex BPs in commercial and scientific domains, where such aspects as initial data acquisition and processing, discovery of BPs with all relevant dependencies, enterprise and life-cycle risk management, selection and execution of the best preservation and redeployment strategies play an important part in providing long-term sustainability of business solutions.

The proposed architecture represents the complete solution for performing preservation of modern BPs and utilises the latest innovations in the area of digital preservation and combines the unique set of knowledge and expertise of all TIMBUS partners.

4. ACKNOWLEDGEMENTS

This publication would not have been possible without the support of TIMBUS project members. The authors wishes to express their gratitude to all TIMBUS members who were abundantly helpful and offered invaluable assistance, support and guidance.

5. REFERENCES

- [1] Access to audiovisual archives. (AXES); CT-2009.4.1 Digital Libraries and Digital Preservation, Project reference: 269980, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/97842_en.html.
- [2] Advanced workflow preservation technologies for enhanced science. (WF4EVER); ICT-2009.4.1 Digital Libraries and Digital Preservation, Project reference: 270192, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/97462_en.html.
- [3] Alliance permanent access to the records of science in europe network. (APARSEN); ICT-2007.4.1 Digital libraries and technology-enhanced learning, Project reference: 269977, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/97472_en.html.
- [4] Archive communities memories. (ARCOMEM); ICT-2009.4.1 Digital Libraries and Digital Preservation, Project reference: 270239, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/97303_en.html.
- [5] Augmented reality supported adaptive and personalised experience in a museum based on processing real-time sensor events. (CULTURA); ICT-2009.4.1 Digital Libraries and Digital Preservation, Project reference: 270318, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/97475_en.html.
- [6] Blogforever. (BLOGFOREVER); ICT-2009.4.1 Digital Libraries and Digital Preservation, Project reference:

- 269963, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/98063_en.html.
- [7] Cultivating understanding and research through adaptivity. (CULTURA); ICT-2009.4.1 Digital Libraries and Digital Preservation, Project reference: 269973, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/97304_en.html.
- [8] Cultural and historical digital libraries dynamically mined from news archives. (PAPYRUS); ICT-2007.4.1 Digital libraries and technology-enhanced learning, Project reference: 215874, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/85544_en.html.
- [9] Cultural heritage experiences through socio-personal interactions and storytelling. (CHESS); ICT-2009.4.1 Digital Libraries and Digital Preservation, Project reference: 270198, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/97182_en.html.
- [10] Cultural heritage of christian ethiopia: Salvation, preservation and research. (ETHIO-SPARE); ERC-SG-SH5 ERC Starting Grant - Cultures and cultural production, Project reference: 240720, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/92358_en.html.
- [11] Digital environment for cultural interfaces; promoting heritage, education and research. (DECIPHER); ICT-2009.4.1 Digital Libraries and Digital Preservation, Project reference: 270001, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/97302_en.html.
- [12] Digital preservation for timeless business processes and services. (TIMBUS); ICT-2009.4.1 Digital Libraries and Digital Preservation, Project reference: 269940, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/99180_en.html.
- [13] Enabling knowledge sustainability usability and recovery for economic value. (ENSURE); ICT-2009.4.1 Digital Libraries and Digital Preservation, Project reference: 270000, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/98002_en.html.
- [14] Five centuries of marriages. (5COFM); ERC-AG-SH6 ERC Advanced Grant - The study of the human past, Project reference: 269796, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/98760_en.html.
- [15] Ground european network for earth science interoperations digital repositories. (GENESI-DR); INFRA-2007-1.2.1 Scientific Digital Repositories, Project reference: 212073, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/92602_en.html.
- [16] Living web archives. (LIWA); ICT-2007.4.1 Digital libraries and technology-enhanced learning, Project reference: 216267, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/85330_en.html.
- [17] Personalised access to cultural heritage spaces. (PATHS); ICT-2009.4.1 Digital Libraries and Digital Preservation, Project reference: 270082, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/97476_en.html.
- [18] Preparing for the construction of the digital research infrastructure for the arts and humanities. (PREPARINGDARIAH); INFRA-2007-2.2-01 Preparatory phase for the projects in the 2006 ESFRI Roadmap, Project reference: 211583, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/88504_en.html.
- [19] Preservation organizations using tools in agent environments. (PROTAGE); ICT-2007.4.1 Digital libraries and technology-enhanced learning, Project reference: 216746, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/85354_en.html.
- [20] Scalable preservation environments. (SCAPE); ICT-2009.4.1 Digital Libraries and Digital Preservation, Project reference: 270137, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/97458_en.html.
- [21] Study of open access publishing. (SOAP); SiS-2008-1.3.1.1 Coordination and support actions on the scientific publishing system in connection with research excellence and dissemination and sharing of knowledge , Project reference: 230220, Retrieved April 18, 2012, from FP7-ICT Website: http://cordis.europa.eu/projects/rcn/91049_en.html.
- [22] G. J. Baida, Z. and B. Omelayenko. A shared service terminology for online service provisioning. In *In Proceedings of the 6th international conference on Electronic commerce (ICEC '04)*, pages 1–10. ACM, New York, NY, USA, 2004.
- [23] B. Birrell, A. and Nelson. Implementing remote procedure calls. *ACM Transactions*, 2(1):39–59, 1984.
- [24] V. K. Cardoso, J. and M. Winkler. Service engineering for the internet of services. *Enterprise Information Systems*, (19):15–27, 2009.
- [25] R. Desisto and B. Pring. Essential saas overview and 2011 guide to saas research. *Gartner*, pages 1–16, 2011.
- [26] T. Erl. Service-oriented architecture: Concepts, technology, and design. *Prentice Hall PTR, Upper Saddle River, NJ, USA.*, 2005.
- [27] S. Kounev and A. Buchmann. *Vedran Kordic (ed.) Petri Net, Theory and Application*, chapter On the Use of Queueing Petri Nets for Modeling and Performance Analysis of Distributed Systems. Advanced Robotic Systems International, I-Tech Education and Publishing, Vienna, Austria, FEB 2007.
- [28] T. Mowbray and W. Ruh. Inside corba - distributed object standards and applications. *Addison-Wesley-Longman*, I-XXIII:1–376, 1988.
- [29] L. Richardson and S. Ruby. Restful web services. (*First ed.*). *O'Reilly*, 2007.
- [30] J. Taylor. From p2p to web services and grids. *Springer, London*, 2005.

Towards a Decision Support Architecture for Digital Preservation of Business Processes

Martin Alexander Neumann
KIT, TECO
Karlsruhe, Germany
mneumann@teco.edu

Goncalo Antunes
INESC ID
Lisbon, Portugal
goncalo.antunes@ist.utl.pt

Hossein Miri
KIT, TECO
Karlsruhe, Germany
miri@teco.edu

Rudolf Mayer
Secure Business Austria
Vienna, Austria
mayer@sba-research.at

John Thomson
Caixa Magica Software
Lisbon, Portugal
john.thomson@caixamagica.pt

Michael Beigl
KIT, TECO
Karlsruhe, Germany
beigl@teco.edu

ABSTRACT

In this paper, we present and address a number of challenges in digital preservation of entire business processes: (1) identifying digital objects a business process depends on (“What to preserve and why?”); (2) identifying significant changes in digital objects (“When to preserve and why?”); (3) determining a re-deployment setting (“What to re-deploy and why?”). After highlighting these challenges, we illustrate some aspects of business processes that are relevant in the context of digital preservation and provide a model to capture their semantics formally. We, then, proceed to present a decision support architecture to address the challenges using the developed model. We, finally, conclude the paper by discussing the applicability of our proposed model and its associated techniques.

Keywords

Digital Preservation, Decision Support, Business Processes

1. INTRODUCTION

Digital preservation research is concerned with providing long-term access to and intelligibility of digital objects, regardless of their complexity. It focuses on preserving digital objects along with their meta-data (or contextual information) required to achieve this goal [10]. In the past, the digital preservation research has been concerned about digital objects which are static in nature, meaning they do not perform active behaviour¹ over time. In digital preservation communities, such as libraries, archives, and museums, this includes text and multimedia documents. Notably, digital objects are generated and interpreted using computational environments [9].

Recent digital preservation research activities have focused on extending established preservation approaches to dynamic digital ob-

jects; referring to those that actively perform behaviour over time. Examples of such dynamic digital objects are video games[14], interactive art[21, 2] and computational environments, such as computational scientific workflows[26]. Furthermore, an increasing amount of static digital objects are being replaced by dynamically generated ones—e.g. dynamic websites, results of e-science experiments, generated meta-data, etc. This content is generated using processes (i.e. computational environments) such as the simplified documents classification process depicted in Figure 1. This means that in order to preserve digital objects in general, the processes that define the context, within which objects are accessed and interpreted, have to be preserved as well.

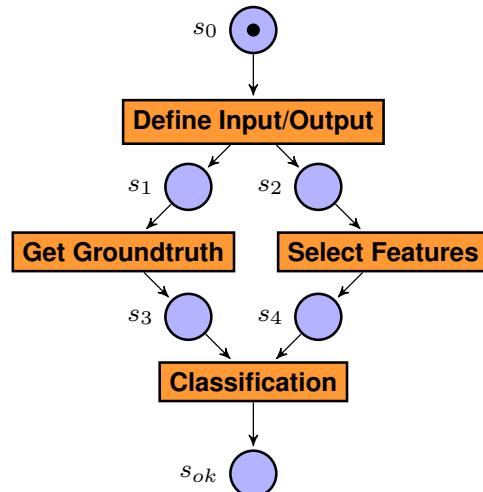


Figure 1: Classification Process to be Preserved

¹Active behaviour describes any externally-visible actions performed by the digital object to interact with its environment. It also refers to any actions performed purely internally which are not externally visible.

To provide long-term continuity in business, we are interested in digital preservation of business processes [12]. But modern business processes form considerably more complex dynamic ecosystems. A process may span many involved legal parties, is interacted with by many people having varying roles, concerns, responsibilities and authorizations, and is supported by a complex distributed service infrastructure. We, therefore, present and address here a number of challenges for the digital preservation of entire business processes that have been identified in a current digital preservation project. The project focuses on time-resilient business processes, and defines

the procedures for the preservation of whole business processes as: (a) preservation planning, (b) preservation execution, and (c) preservation re-deployment (also referred to as exhumation of a preserved process). In the context of these procedures, the relevant challenges include:

What to preserve and why? During preservation planning, we identify *what* digital objects a business process depends on and *why*.

When to preserve and why? During preservation planning, we identify the relevant differences in digital objects to determine *when* to preserve a business process and *why*.

What to re-deploy and why? Before re-deployment, we identify *what* are suitable re-deployment settings, in terms of *what* preserved digital objects will be re-deployed in *what* re-deployment environments² and *why*.

In Section 1.1, we discuss the context of business processes relevant to digital preservation and how to model it. In Section 1.2 we discuss how to establish decision support for digital preservation activities based on these models.

In Section 1.3, we point out three reasoning tasks in the context of preservation planning, execution and re-deployment for business processes. In order to define the scope of this paper, we only focus on these three tasks (which are closely related and involve the entire preservation process). Section 2 illustrates the proposed model that has been developed for the digital preservation of business processes (which will be further revised in future). This model captures knowledge which is generally relevant to digital preservation of business processes, based on a set of representative use-cases and an enterprise modelling framework.

In Sections 3 and 4, we explain how we address the reasoning tasks based on our model and a proposed decision support architecture. We also analyze the computational complexity of our three proposed approaches. Finally, we discuss the applicability of our approach to preservation of business processes, and conclude in Section 5.

1.1 Relevant Context of Business Processes

We argue that there are many aspects in the context of a business process that have to be taken into account during preservation planning and execution, to ensure successful re-deployment of that process. We consider *successful re-deployment* as the ability to re-run a preserved process which behaves in the same way as the original one³ [1]. Additionally, we argue that, in the context of a business process, (1) there are abstract (coarse-granular) aspects which are relevant to the entire domain of process preservation, and (2) there are more specific aspects (fine-granular) which are relevant to sub-domains of process preservation, e.g. the *class of scientific processes* or an

²An adjustable part of a re-deployment environment may be adapted during the re-deployment procedure to accommodate for the process-specific situation established by the preserved digital objects and parts of a re-deployment environment which are fixed.

³It behaves equivalent according to an equivalence notion, such as trace equivalence[24], and equivalent in terms of relevant modalities, such as causality and time. Both aspects are determined by the requirements of process preservation in general, but also by the requirements of preserving the process in focus.

individual scientific experiment, which may identify further relevant aspects. For example, at the most coarse-granular level, we have identified the following abstract categories of aspects as being relevant to the entire domain of business process preservation: (1) processes, (2) preservation requirements, (3) services, (4) software, (5) hardware, (6) data (7) licenses, (8) authorizations, and (9) people. The elements of these categories combine to form a complex inter-dependent network of different types of classes, individuals, relations and rules—they form an upper ontology capturing the knowledge relevant to business process preservation in general. This ontology may be lowered to sub-domain- or even process-specific ones to capture the knowledge relevant to the respective sub-domain.

In terms of decision support for preservation activities, there is an issue of these aspects forming large networks. Conceptually, we can use these networks of aspects to assist us in drawing conclusions from them, as illustrated in Section 1.3. However, the networks' complexities could hinder digital preservation engineers from sketching them on a blackboard and manually drawing conclusions. If we model these aspects and their inter-relations semantically adequately, we can support planning, execution and re-deployment activities using reasoning on these models. *Semantically adequately modelled* means that the model captures the semantics of the business process and its context in such a way that is suitable for automatically drawing conclusions of practical use for process preservation. The practical suitability of our model and results derived by reasoning on it have to be experimentally evaluated.

In this paper, the *context relevant to digital preservation* refers to any information that a designated user community requires to comprehend the preserved digital objects properly—i.e. intelligibility of digital objects to a designated group of people at some future point in time [10]. There are several models in the literature that capture information on context relevant to the digital preservation of digital objects. According to the Open Archival Information Systems (OAIS) Reference Model, this information is separated into *representation information* (structure and semantic information, and a representation network) and *preservation description information* (reference, context, provenance, and fixity information) [17].

Examples of models (and related formats) are: (1) the METS and OAI-ORE formats for packaging and exchanging of digital resources; (2) FRBR[16] to model information realization and versioning problems in libraries; (3) Dublin Core, MODS and MARC to record bibliographic information; (4) the ABC Ontology[19], the Open Provenance Model[22], the PROV Data Model[4], the SHAMAN Context Model[6], and the PREMIS Data Dictionary for Preservation Metadata[25] capture provenance information to model life-cycles of digital objects in and/or outside of digital archives; (5) CIDOC CRM[22] to integrate heterogeneous cultural heritage information; and (6) representation information networks[17] to structure representation information.

These models provide means for modelling OAIS-relevant information on digital objects with different focus and varying levels of detail. They are concerned about structural generic semantic aspects of digital objects, and about processes in the context of digital objects. But they do not yet characterize “behavioural aspects” of (dynamic) digital objects themselves. In addition, they do not yet focus on semantic aspects specifically relevant in the context of business processes or workflows.

From our perspective, executional aspects are relevant, because we have to model systems which are complex objects on the one hand (as business processes have a compositional structure of inter-related parts), and those which perform actions (behave) on the other hand. Thus, in addition to a structural and generic semantics notion and model, we need a notion and model of behaviour which is adequately applicable to digital preservation of business processes. As stated before, this notion and model of behaviour has to accomplish the above goal of enabling successful re-deployment of a preserved process. As a consequence, we extend the interpretation of the term *digital preservation relevant context* to: information that a designated user community requires to comprehend archived digital objects properly, as well as information that a designated user community requires to verify the execution of a re-deployed behavioural system. We also propose a novel modelling approach for the digital preservation of business processes that captures relevant structural, semantic, and behavioural aspects, to enable successful re-deployment of a preserved process. However, as mentioned above, whether the modelling approach achieves this goal has yet to be evaluated in representative case studies of whole process preservation.

To foster preservation of computational scientific workflows, models for context and behaviour of such processes are proposed in [26, 13]. Context is modelled as sets of required services and data in [26], and [13] proposes a notion of process behaviour which seems equivalent to condition-event structures (which are revisited in Section 2 and we promote too since [23]). To build on this research, in this paper, we extend our notion of process behaviour by time and propose a flexible context modelling approach.

1.2 Decision Support for Digital Preservation

As mentioned before, the introduced models used for capturing the context relevant to digital objects focus on their respective domains which they model to a certain level of detail and at a certain level formality. From a knowledge representation perspective, they all are based on individual domain-specific ontologies; i.e. in general, to model digital objects and information about their context, the ontologies provide relevant: (1) classes, (2) instances of these classes, (3) relations between these classes and instances, and (4) additional rule-like statements on classes, individuals and relations.

Enabling tractable automated reasoning on these models requires them to be based on an adequately expressive and decidable language which sound and complete inference mechanisms can operate on. This provides the ability to provide explainable and correct answers to any expressible decision problem or query on the models in feasible time. The required level of formality is provided by some of the covered models. For example, the Open Provenance Model, the PROV Data Model, the ABC Ontology, the PREMIS Data Dictionary, and CIDOC CRM have been implemented in the Web Ontology Language 2 DL (OWL 2 DL)[15] (or subsumed language fragments).

Besides capturing behavioural aspects, our modelling approach captures the introduced structural and semantic aspects. Both are modelled on a “semantically rich” (i.e. formal and detailed) level, based on an ontology language in general. This has two advantages: (1) automated reasoners that assist during preservation planning, execution, and re-deployment can directly operate on the knowledge maintained along with a preserved digital object; and (2) the knowledge kept with a preserved digital object can even be specific to this object, which means that the model is specific to the preserved business process. A reasoner would, then, directly be able

to draw conclusions from it without having to combine the knowledge kept with the digital object with the background knowledge kept inside the reasoner itself. Combining both would be necessary, if the reasoner would bring in some knowledge in addition to the knowledge kept with a digital object. In this case, both knowledge bases are in danger of contradicting each other and, therefore, hard to combine [7]—in particular, if both knowledge bases originate from different contexts, such as points in time or user communities. This implies another positive of our approach: in general, reasoners do not have to be sub-domain- or process-specifically adapted and are thus time-resilient.

As already mentioned, we promote the use of an ontology to model the information and knowledge on digital objects, and also to design object-specific models to accommodate for specific digital preservation requirements of an object. For example, in one scenario it might be sufficient for re-deployment of a business process if the requirements stipulate causal trace-equivalent behaviour after re-deployment. However, in the case of a scientific experiment, causality and exact timing are likely to be very relevant. Therefore, if we would like to assist preservation planning in answering the question “what to preserve and why?” for both processes, there is no generic strategy to answer it. For the first process, it would be sufficient to only preserve technical requirements down to the operating systems which in this example are known to provide a runtime environment that preserves causality. In the case of the second process, we might need to preserve technical requirements down to the hardware, which is assumed to provide cycle-time accurate timing. Therefore here, we need two different strategies (or policies) to determine which parts of the business processes are required to be captured. As the strategy is specific to the digital object in focus, it must be kept with the object itself and not the reasoner.

We envision that many digital preservation related questions are specific to digital objects, analogous to the illustrated example. Answering these digital preservation questions depends on the context (or situation). Therefore, we argue that it is important to provide the ability to capture object-specific knowledge for their digital preservation, in particular for business processes. This would improve the understanding of preserved digital objects without the need for background knowledge, and also enable generic reasoning mechanisms to act on the preserved digital object only, to assist in preservation activities, such as planning, execution, and re-deployment.

The digital preservation research has already implemented decision support approaches. The most recent one is Plato[3]. In contrast to our methodology, Plato focuses on digital objects which are static in nature, and as such do not perform active behaviour over time; e.g. text documents and images. Plato provides a reasoning framework for identifying relevant actions to preserve a digital object. In general, this idea complements the approach pursued in this paper, as we do not discuss the question of “how to preserve a digital object?”. And, as we are concerned about dynamic digital objects, Plato’s applicability to this domain is a relevant future aspect.

To achieve its goal, Plato (1) defines generic features of digital objects, such as the presence of intellectual property rights issues; (2) defines more specific features of classes of digital objects, such as compression characteristics of image formats; (3) devises methods to extract these features from digital objects, such as by using tools or performing manual experiments; and (4) proposes a method to conclude optimal preservation actions from the features of a

digital object. This methodology is in line with our vision and requirement of being able to draw conclusions from the model of a digital object only. To provide this, a generic mechanism is proposed that calculates and compares the *utilities* of preservation actions on a unified scale, whereby the feature extraction techniques of a digital object are responsible for providing a strategy to map their outputs onto this scale.

1.3 Process Preservation Challenges

In order to be correctly rendered, a digital object needs a technological context resulting from the combination of specific hardware and software. Moreover, in order to be correctly understood by humans, the organizational, business, and social contexts surrounding the object are also needed. The Digital Preservation Europe Research Roadmap, published in 2007, defines the context of a digital object as the “representation of known properties associated with and the operations that have been carried out on it”[11]. On the one hand, these properties might include information about the technology used, but on the other hand they might consist of legal requirements, existing knowledge, and user requirements. The operations performed on an object might even include the processes that originated the object itself.

The determination of the relevant context of a digital object becomes even more challenging if complex digital objects such as workflow or business process specifications are considered. Those types of objects are dependent on a highly complex and distributed technical infrastructure hosted in complex and diverse organizational settings, sometimes involving multiple organizations. This creates a complex dependency network involving the object and other complex objects on which its correct rendering and understanding depends. However, not all context might be relevant for being able to preserve and successfully re-deploy a process in the future. Some of the context might not even be available at all—for example, if the details of some external services are not accessible. In general, a selective approach for determining the context of a process should be pursued, which enables to select the partial context which is use case-specifically required for preservation of a process. Otherwise, it might lead to resource waste, and might even cause the costs of preservation to surpass its potential benefits. In that sense, the first preservation challenge faced when dealing with the preservation of business processes is “what to preserve and why?”.

After the identification of the relevant contextual information, it becomes necessary to determine how to approach the capturing and preservation of the process and relevant context. In other words, it is important to determine what preservation actions should be performed. As introduced, this issue has so far been addressed by Plato. It is assumed that surpassing this challenge will result in the successful execution of the preservation actions that will allow the process and its relevant context to be preserved.

Furthermore, as a process and its context have to be captured at a determined point in time during preservation, it becomes crucial to monitor the original process to detect any changes in process behaviour. Since those changes are potentially relevant to capture to preserve the most recently working version of a process, another preservation challenge being faced is “when to preserve and why?”. Facing this challenge successfully will involve having several snapshots of a process and its relevant context documenting the main events happening during its life-cycle.

Challenges are also faced during the re-deployment of a preserved

process. Since digital preservation concerns the long-term, it is highly probable that the original deployment setting is partly or not available at all. The preserved context model provides indicators to what are suitable re-deployment settings for the preserved processes. The re-deployed environment might need adaptation during the re-deployment procedure in order to re-establish any situation of interest. In general, an optimizing approach for determining re-deployment settings should be pursued, to minimize re-deployment efforts and therefore associated costs. Hence, a challenge that must also be faced in the re-deployment of business processes includes knowing “what to re-deploy and why?”.

After the identification of the re-deployment setting, it becomes necessary to determine how to approach the re-deployment itself. Thus, it is crucial to determine what re-deployment actions should be performed. This issue is, again, analogous to what is already being addressed by Plato. And it will be surpassed if the re-deployment of the process and environment allow for the correct re-execution of the process. This is an issue we are trying to resolve by comparing a re-deployed process to its original process based on the comparison of the outputs produced by them, as presented in [20].

2. CONTEXT MODEL

Our *context model* describes business processes and their context, both of which are scoped to aspects relevant to the digital preservation of the processes. The context model is a formal ontology that can be instantiated, or specialized, to model individual digital preservation settings (which involve concrete business processes and their context). The instantiation of the model involves the definition of classes, individuals, relations, and statements which are specific to the digital preservation setting. This provides the ability to model processes and their digital preservation-relevant context in a semantically rich way, as motivated in Section 1. To specify our ontology and scope it to the domain of digital preservation of business processes, we have investigated which classes, individuals, relations, and logical statements apply to the entire domain of digital preservation of business processes. The design methodology (middle-out approach) and preliminary details on the contents of our ontology are presented in [20].

Furthermore, as introduced in [23], we have identified *condition-event structures* (or 1-safe petri nets) as being an adequate notion for modelling the structure and causal behaviour of business processes. It is an approach for design and *efficient* verification which clearly formulates causal behaviour of concurrent systems [8]. To be able to additionally model temporal behaviour of business processes, as required in this work, we extend our notion to *time condition-event structures*. This approach allows to model *causal and temporal behaviour* of concurrent processes for design and verification.

A model $\mathcal{M} := \langle \mathcal{B}, \mathcal{C} \rangle$ consists of a set of business processes \mathcal{B} and a context \mathcal{C} . A condition-event structure $\mathcal{N}^{c/e} := \langle \mathcal{P}, \mathcal{T}, \mathcal{F}, m_0 \rangle$ consists of a set of places \mathcal{P} encoding *conditions* and a set \mathcal{T} of transitions encoding *events*, where $\mathcal{F} \subseteq (\mathcal{P} \times \mathcal{T}) \cup (\mathcal{T} \times \mathcal{P})$ is the set of edges of the net and m_0 is the *initial marking*. Here, a function $m_i : \mathcal{P} \rightarrow \{0, 1\}$ is called a *marking*. A transition t is activated (“may fire”) in a marking m_i iff for all p holds: (1) if $(t, p) \in \mathcal{F}$ then $m_i(p) = 0$, and (2) if $(p, t) \in \mathcal{F}$ then $m_i(p) = 1$. A sequence of “fired” transitions $t_i \rightarrow \dots \rightarrow t_j$ is called a *trace*.

A time condition-event structure $\mathcal{N}^{t,c/e} := \langle \mathcal{P}, \mathcal{T}, \mathcal{F}, m_0, l \rangle$ consists of a condition-event structure $\langle \mathcal{P}, \mathcal{T}, \mathcal{F}, m_0 \rangle$ and a *time labelling function* $l : \mathcal{T} \rightarrow \mathbb{N}_{\geq 0} \times \mathbb{N}_{\geq 0} \cup \{\infty\}$ whereby for all

$t = (t_i^{\circ}, t_i^{\bullet})$ holds: $t_i^{\circ} \leq t_i^{\bullet}$ and $t_i^{\bullet} < \infty$. All t_i° are called *earliest firing times* and all t_i^{\bullet} are called *latest firing times*. A transition “may fire” the earliest at its t_i° and “has to fire” the latest at its t_i^{\bullet} since its activation. Furthermore, $j_i : \mathcal{T} \rightarrow \mathbb{N}_{\geq 0} \cup \{\phi\}$ is a *clock function* that gives the time which has elapsed since a transition t has been activated. In consequence, for all t_i holds: $j_j(t_i) \geq t_i^{\circ}$ and $j_j(t_i) \leq t_i^{\bullet}$. A sequence of time-annotated “fired” transitions $(t_i, j_i) \rightarrow \dots \rightarrow (t_j, j_j)$ is called a *time trace*.

Now, the set of business processes \mathcal{B} in our model can be defined as a set of time condition-event structures: $\mathcal{N}_i^{t,c/e} \in \mathcal{B}$. Furthermore, the context $\mathcal{C} := \langle \mathcal{E}, \mathcal{R}, \mathcal{S} \rangle$ consists of a set of classes \mathcal{E} , a set of relations \mathcal{R} and, a set of logical statements \mathcal{S} . Each class $e_i := \{i_0 \dots i_n\}$ is a set of individuals i_j . Each relation $r_i \subseteq (\mathcal{T} \times \mathcal{E}) \cup (\mathcal{E} \times \mathcal{E})$ relates transitions (i.e. events) to classes, and classes to classes. Each logical statement s_i is a horn-formula in first-order logic[18] whereby its predicates are restricted to the relations in \mathcal{E} and \mathcal{R} .

3. ADDRESSING THE CHALLENGES

Figure 2 presents our proposed architecture to provide decision support in terms of the highlighted challenges. In a concrete digital preservation setting, the context model (1) introduced in Section 2 is firstly fed into the *Model Builder* to create a specialized instance of the model—it ingests our ontology which is specific to the entire domain of process preservation to create an instance of it specific to the process. Secondly, to create this instance, relevant knowledge from knowledge bases⁴ (2—such as data formats and software licenses) and process-specific details (3—such as process-specific preservation requirements, and involved software and hardware) are added to the ontology by the *Model Builder*. The process-specific details may either be automatically extracted from a business process (e.g. software and hardware) or manually input by digital preservation engineers (e.g. preservation requirements).

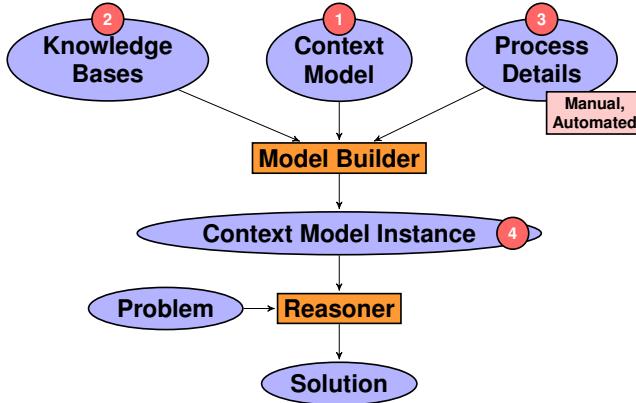


Figure 2: Decision Support Architecture

The produced model (4) captures all the knowledge relevant to the digital preservation of the process in focus and will accompany the process during its entire life-cycle in a preservation archive. Furthermore, the model contains the knowledge required to provide decision support to the three presented preservation challenges, as will be illustrated in the following sections. In general, as our model is based on individuals (objects), classes (unary relations), binary

⁴The knowledge bases conceptually are part of the ingested context model, but are kept separate from it in our implementation.

relations and horn formula in the two-variable fragment of first-order logic, the produced model can be handed over to various types of semantic reasoners (such as “off-the-shelf” description logic or first-order logic reasoners) to conclude solutions from given problems based on the given model only.

3.1 Objects to Preserve

As outlined earlier, answering the question of “what to preserve and why?” can be reduced to establishing a notion of what is *required* by a process to be preserved and successfully re-deployed. This notion is determined by preservation requirements which are relevant to the entire domain of process preservation, and more specific requirements which are relevant to sub-domains of process preservation. For example, as illustrated earlier, each process imposes individual requirements on its causality and timing equivalence. Therefore, this notion is specific to the process and the digital preservation setting⁵ (called *process-specific* in the following).

In general and in our ontology, there are several ways to model *what is required* by a process. One approach is to explicitly model a *requires relation*. For example, we could say that “a program requires an operating system, which requires a machine, which requires an operator”. This is a semantically rather limited notion, and there is no need for our idea of a “semantically rich” context model. But this approach does not provide a process-specific notion of what is required. If we capture a model of only *requires relations*, for example, of all software and hardware components involved in a process, we cannot tell what components are “really necessary” for successful re-deployment without inspecting the model and deleting information. This approach is likely to lose information relevant to yet unknown re-deployment settings.

Another approach would be to implicitly model a *requires relation* by declaring other relations, such as *runsOn*, *isInstalledOn*, *isOperatedBy* of being a subtype of the *requires relation*. Based on this, we could process-specifically select which relations determine *what is required*. For example, we could model that “a program isInstalledOn an operating system, which runsOn a machine, which isOperatedBy an operator” and conclude that all four individuals are required by our process. But this is still semantically rather limited, as we still could not process-specifically distinguish, for example, “really necessary” software and hardware components from “not really necessary” ones.

Therefore, we argue that a more expressive approach is required which provides a more complex notion of *what is required*, and we propose horn formula in the two-variable fragment of first-order logic to express this process-specifically in our ontology. It allows to express that all objects that satisfy a complex statement are required. For example, “it is only necessary to preserve an operating system if it is proprietarily licensed”. We are in the process of implementing this approach using OWL 2 DL and the Pellet reasoner[27]. Based on this, all the problems presented to our reasoning engine are decidable, although the employed language exposes a worst-case computational complexity in reasoning of N2EXPTIME[5]. Our future efforts will determine which language fragments are required in the process preservation practice to improve on the complexity and whether it is a computationally tractable approach.

⁵In this context, the *setting* particularly refers to the temporal preservation horizon which determines setting-specific aspects such as available technologies and relevant user communities of the future.

3.2 Events to Preserve

As mentioned earlier and discussed in [23], answering the question of “when to preserve and why?” can be reduced to establishing a notion of what is the *difference* between the process now and when it has been preserved the last time. If this difference exceeds some level of relevance, then a new trigger to preservation execution is determined. Again, this notion of *what a relevant difference in what modalities is*, is process-specific, as each process imposes individual requirements on its causality and timing equivalence.

We propose a notion of *trace equivalence* to detect *relevant differences* in causality and timing behaviour of a process at two different times. Our idea is based on the detection of relevant differences in the execution traces of processes under equivalent contextual conditions (regarding their interaction with the environment, such as values of inputs). Based on the *traces* and *time traces* of processes that are defined in our model (in Section 2), we can compare traces stored in two models with each other. Comparing any two traces requires that they have been taken under equivalent contextual conditions—they are called *comparable traces* in the following. We propose a *process-agnostic* notion of difference in the qualitative order of events, and a *process-specific* notion of difference in the quantitative order of events.

Regarding the qualitative difference notion, any change in the qualitative order of events between two comparable traces marks a *relevant difference*. Regarding the quantitative difference notion, *deviations* of an event’s timing (in a *time trace*) from its time interval⁶ marks a trace which deviates from its process specification. Incorporating the process (of which the trace has been taken) is important in this case, as the quantitative difference notion is process-specific. Two *comparable time traces differ relevantly* from each other, if and only if one of them deviates from the timing interval specification and the other one does not. Each process defines an individual interval of expected timing values for each event, as defined in our model in Section 2. These individual interval information can be either given by expert knowledge or by profiling a process.

The causal and timing behaviour of a process, during its execution under specific contextual conditions, is given by one *time trace* in our model. If we want to capture the behaviour of a process under varying contextual conditions, we need to capture (in our model) a set of *time traces*, along with their contextual conditions. To compare two processes, we compare their trace sets. The trace sets have to have been taken under the same varying conditions. Each two traces that have been taken under the equivalent conditions have to be compared with each other. If this fails on at least one set of two traces, a *relevant difference* has been identified. When this approach is applied to monitoring of a process which is to be preserved, the identified relevant difference represents a trigger (“when to preserve and why?”) to preservation of the process.

Analogously to the previous challenge presented in Section 3.1, we are in the process of implementing this approach using a tractable fragment of OWL 2 DL and the Pellet reasoner.

3.3 Objects to Re-Deploy

Although it seems analogous, answering the question of “what to re-deploy and why?” is considerably more complex than the earlier discussed question of “what to preserve and why?”. In

⁶Refers to the time interval specification of the event in the *time condition-event structure* of the process (in our model).

addition to the preserved process, we have to take into consideration the environment we are going to re-deploy the process into. The re-deployment environment will consist of a *fixed* and a *flexible part*. This means that there will be an unchangeable (or constrained) part in the re-deployment environment, for example, some machines in a data center, and a changeable (or un-constrained) part of the environment, for example, the possibility of selecting an alternative operating system running on these machines in the data center. We reduce answering the question “what to re-deploy and why?” to a notion of what is *required* to re-deploy a preserved process. Again, this notion is process-specific, even more than in our previous challenges as the re-deployment environment takes a major role in our reasoning problem.

In reasoning, we have to take three instances of our context model into account, which have to be determined first: a model of the *preserved process*, a model of the *constrained environment*, and a model of the *un-constrained environment*. Afterwards, we will determine all *feasible re-deployment alternatives* and pick an *optimal one*. This is performed by identifying the *difference* between the preserved process and the constrained environment in more detail. There are four possible outcomes of this evaluation:

None The constrained environment is identical to the environment when the process has been preserved. Therefore, combining their models does not introduce inconsistencies into our ontology, and neither our process, nor the environment have to be adapted to re-deploy.

Overlap The preserved process and the constrained environment *overlap*. This means that their combined model contains overlapping sub-graphs which address the same issue, meaning which are not allowed to overlap and therefore introduce inconsistencies into the ontology. For example, two different operating systems on the machines in the data center.

Gap There is a *gap* between the preserved process and the constrained environment. This means that their combined model contain sub-graphs which are disconnected from each other although they need to be connected, meaning the disconnected sub-graphs introduce inconsistencies into the ontology too. For example, if none of the models cover operating systems.

Both The preserved process and the constrained environment partially *overlap* at one to many points and partially have one to many *gaps* between each other.

After the situation has been sorted out thoroughly, and if we have determined that we cannot immediately re-deploy, we continue in a second step to determine all feasible re-deployment alternatives. This is based on the models of the preserved process, and both environment models (constrained and un-constrained). The reasoner applies the following strategies in solving any gaps or overlaps:

Overlap In case of an overlap between the models of the preserved process and the constrained environment, the reasoner will take parts out of the model of the preserved process to find options that eliminate the inconsistency from our ontology. This may mean that the reasoner takes larger parts out of the model than the actual overlap, which are filled by parts from the model of the un-constrained environment.

Gap In case of a gap between the models, the reasoner uses the model of the un-constrained environment to find all options to fill this gap and thus eliminate the inconsistency from the model. This may even mean that the reasoner has to take out parts from the model of the preserved process.

Afterwards, all alternatives are ranked to conclude the optimal re-deployment alternative. We are in the process of implementing this reasoning procedure based on linear optimizers, specifically the APT-PBO solver[28], which allows us to determine many feasible re-deployment alternatives and rank them according to a process-specific cost function. APT-PBO is different from other similar solvers in that it acts as an interactive system and as such the proposed solutions can be navigated and further decisions taken that is likely to be important in the re-deployment scenario.

An illustrative example of a technical scenario is having a preserved software library (used by a business application) that will not work with the re-deployment environment. The library may have had a known security flaw meaning that in a re-deployment environment it would have to be updated to a version that included the security fix. Another possible issue could be that the library cannot be used because of licensing issues or doesn't work in combination with some other system that is in place in the new environment. The reasoner would then, based on the context models, try to determine feasible alternatives to the library to update it and rank them according to criteria. This procedure involves the reasoner trying to determine what else would be affected by updating the library. If other software is affected by the update, this could additionally be notified to an digital preservation engineer and then either a more updated version can be installed or a manually-proposed alternative be applied which fulfils the requirements.

4. DECISION SUPPORT WORKFLOW

In [20], we present in detail a classification process which is also sketched in Figure 1. The process builds a music genre classifier based on features extracted from given training data, and afterwards classifies given input data based on features extracted from them. Notably, the process involves a variety of free and proprietary data formats, such as HTML and MP3, and external services, e.g. for feature extraction. We have modelled its behaviour and required formats, software, hardware and licenses in a context model instance. Based on this, we informally illustrate here the application of our proposed procedure to provide decision support to the challenge of “what to preserve and why?” on this model instance.

Instantiate Context Model *The first step in the decision support workflow is to populate the context model semi-automatically using extraction tools and expert knowledge of digital preservation engineers.*

For example, we have extracted a process model from the employed workflow engine which yields the process' behaviour and its external service dependencies. Furthermore, we have extracted a directed graph of software dependencies of the workflow engine from the software package repository of the operating system.

Specify Requirements *Next, the specific requirements of our digital preservation setting to evaluate the question of “what to*

preserve and why?” have to be specified. This covers conditions which are sufficient to be satisfied by an individual such that it has to be preserved. And this covers conditions which are required to be satisfied by an individual such that it can be preserved.

In this example, we follow a straight-forward approach in specifying whether an individual has to be preserved. Analogously to representation information networks[17], we specify dependencies explicitly by introducing a transitive relation called “requires” which subsumes all other relations in the context model instance. Now, we declare that “being (transitively) required by the process” is sufficient for an individual for having to be preserved.

Furthermore, we assume that we are required to preserve for at least 10 years (i.e. long-term). The knowledge modelled in the context model yields, for example, that required software must not depend on external services. Software individuals can only be preserved if they satisfy this requirement. But for our process we relax this by allowing *feature extractors* to be preservable if they exchange data in a standardized format, such as ARFF.

Specialize Context Model *Now, to provide this relaxation of the digital preservation requirements, the context model has to be inspected and its classes and relations specialized to process-specific needs. At this point, the workflow becomes iteratively, as in the next step the model has to be re-instantiated to populate the specialization appropriately.*

For example, we have added the concept of *feature extractors* (a specialization of external services), which is a relevant concept of our process to reason about its preservability.

Evaluate Results *And finally, our proposed reasoning engine is employed to determine (1) if these requirements can be satisfied, and (2) what sub-graphs of the context model instance satisfy them.*

Without having specified that feature extractors are preservable, our procedure would conclude that the desired long-term preservation cannot be performed—yielding the non-preservable external feature extraction service as the reason. After expert consultation, we have relaxed this requirement, which yields at least one preservable sub-graph of our context model instance.

5. CONCLUSION

In this paper, we have motivated the necessity for digital preservation research on dynamic digital objects, such as processes generating (a) dynamic websites, (b) results in e-science experiments, and (c) metadata. Based on this, we have illustrated three challenges in decision making that span the procedures linked with digital preservation of business processes (planning, execution and re-deployment). These challenges have been identified in the context of a digital preservation project that focuses on time-resilient business processes. The challenges were: (1) identifying digital objects a business process depends on; (2) identifying significant changes in those objects; and (3) determining suitable re-deployment settings. As motivated earlier, due to the complexity of the tasks at hand and its inherently associated efforts, providing techniques in solving them using decision support tools will ease the duties of involved stakeholders.

In previous work, we have already presented ideas to partially address the first and second challenges, and we have outlined their application in a case study, a scientific workflow. A context model instance is semi-automatically generated and a method for verifying the workflow's behaviour after re-deployment is presented in [20]. In this paper we have extended this work by (a) devising a procedure for determining "what to preserve and why?" from a given context model instance, and (b) by specifying an equivalence notion on *time traces* to detect relevant changes in process behaviour on a generic base. Furthermore, in [23], we propose an approach to monitoring of business processes to trigger their digital preservation and verifying their causal behaviour. Here, we have extended this notion to enable verification of causal and temporal behaviour of processes.

In addition, in this paper, we have presented an architecture to assist in the decision making of the preservation procedures in general. The architecture has been based on a knowledge representation technique specifically tailored to process preservation, called the *context model*. We have, also, presented how we are addressing the identified challenges using the architecture and reasoners applicable to our model—in general, logic-based reasoning engines (Pellet and APT-PBO) being applied. In [12] we present the integration of the model and several instances of our proposed architecture (which address the challenges) into an architecture for digital preservation of entire business processes. Our future efforts are focused on implementing and evaluating the covered modules.

6. ACKNOWLEDGEMENTS

The authors would like to acknowledge the funding by the European Commission under the ICT project "TIMBUS" (Project No. 269940, FP7-ICT-2009-6) within the 7th Framework Programme.

7. REFERENCES

- [1] J. Barateiro, D. Draws, M. A. Neumann, and S. Strodl. Digital Preservation Challenges on Software Life Cycle. pages 487–490, 2012.
- [2] C. Becker, G. Kolar, J. Kueng, and A. Rauber. Preserving Interactive Multimedia Art: A Case Study in Preservation Planning. In *Proceedings of the 10th International Conference on Asian Digital Libraries*, ICADL'07, pages 257–266, Berlin, Heidelberg, 2007. Springer-Verlag.
- [3] C. Becker and A. Rauber. Decision criteria in digital preservation: What to measure and how. *Journal of the American Society for Information Science and Technology*, 62(6):1009–1028, June 2011.
- [4] K. Belhajame, H. Deus, D. Garijo, G. Klyne, P. Missier, S. Soiland-Reyes, and S. Zednik. *PROV Model Primer*. World Wide Web Consortium, 2012.
- [5] R. Bembenik, L. Skonieczny, H. Rybiński, and M. Niegzodka. *Intelligent Tools for Building a Scientific Information Platform*. Studies in Computational Intelligence. Springer-Verlag, 2012.
- [6] H. Brocks, A. Kranstedt, G. Jäschke, and M. Hemmje. Modeling Context for Digital Preservation. In *Smart Information and Knowledge Management*, volume 260 of *Studies in Computational Intelligence*, pages 197–226. Springer, 2010.
- [7] A. Cali, T. Lukasiewicz, L. Predoiu, and H. Stuckenschmidt. Tightly Coupled Probabilistic Description Logic Programs for the Semantic Web. *Journal on Data Semantics*, pages 95–130, 2009.
- [8] A. Cheng, J. Esparza, and J. Palsberg. Complexity results for 1-safe nets. *Theoretical Computer Science*, 147(1-2):117–136, Aug. 1995.
- [9] C. Chou, A. Dappert, J. Delve, and S. Peyrard. Describing Digital Object Environments in PREMIS. In *Proceedings of the 9th International Conference on Preservation of Digital Objects*, iPRES 2012, 2012. (to appear).
- [10] M. Day. Metadata for Digital Preservation: A Review of Recent Developments. In *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '01, pages 161–172, London, UK, UK, 2001. Springer-Verlag.
- [11] DigitalPreservationEurope Partners. DPE Digital Preservation Research Roadmap. Public Deliverable D7.2, DPE, 2007.
- [12] M. Galushka, P. Taylor, W. Gilani, J. Thomson, S. Strodl, and M. A. Neumann. Digital Preservation Of Business Processes with TIMBUS Architecture. In *Proceedings of the 9th International Conference on Preservation of Digital Objects*, iPRES 2012. (to appear).
- [13] D. Garijo and Y. Gil. A new approach for publishing workflows: abstractions, standards, and linked data. In *Proceedings of the 6th Workshop on workflows in Support of large-scale science*, WORKS '11, pages 47–56, New York, NY, USA, 2011. ACM.
- [14] M. Guttenbrunner, C. Becker, and A. Rauber. Keeping the Game Alive: Evaluating Strategies for the Preservation of Console Video Games. *International Journal of Digital Curation*, 5(1), 2010.
- [15] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph. *OWL 2 Web Ontology Language: Primer*. World Wide Web Consortium, 2009.
- [16] International Federation of Library Associations. Functional Requirements for Bibliographic Records - Final Report. Technical report, IFLA, 1998.
- [17] International Organization For Standardization. OAIS: Open Archival Information System – Reference Model. 2003. ISO 14721:2003.
- [18] R. A. Kowalski. Predicate Logic as Programming Language. In *IFIP Congress*, pages 569–574, 1974.
- [19] C. Lagoza and J. Hunter. The ABC Ontology and Model. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*, pages 160–176. National Institute of Informatics, Tokyo, Japan, 2001.
- [20] R. Mayer, A. Rauber, M. A. Neumann, J. Thomson, and G. Antunes. Preserving Scientific Processes from Design to Publications. In *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries*, TPDL 2012. Springer-Verlag, 2012.
- [21] A. McHugh, L. Konstantelos, and M. Barr. Reflections on Preserving the State of New Media Art. In *Proceedings of the Archiving Conference*, 2010.
- [22] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, and J. V. den Bussche. The Open Provenance Model core specification (v1.1). *Future Gener. Comput. Syst.*, 27(6):743–756, June 2011.
- [23] M. A. Neumann, T. Riedel, P. Taylor, H. R. Schmidtke, and M. Beigl. Monitoring for Digital Preservation of Processes. In *Proceedings of the 7th international and interdisciplinary conference on Modeling and using context*, CONTEXT '11, pages 214–220, Berlin, Heidelberg, 2011. Springer-Verlag.
- [24] L. Pomello, G. Rozenberg, and C. Simone. A survey of equivalence notions for net based systems. In *Advances in Petri Nets 1992*, pages 410–472. Springer-Verlag, London, UK, UK, 1992.
- [25] Premis Editorial Committee. Data Dictionary for Preservation Metadata: PREMIS version 2.0. (March), 2008.
- [26] D. D. Roure, K. Belhajame, P. Missier, J. M. Gómez-Pérez, R. Palma, J. E. Ruiz, K. Hettne, M. Roos, G. Klyne, and C. Goble. Towards the Preservation of Scientific Workflows. In *Proceedings of the 8th International Conference on Preservation of Digital Objects*, iPRES 2011.
- [27] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz. Pellet: A practical OWL-DL reasoner. *Web Semantics*, 5(2):51–53, June 2007.
- [28] P. Trezentos, I. Lynce, and A. L. Oliveira. Apt-pbo: solving the software dependency problem using pseudo-boolean optimization. In *Proceedings of the IEEE/ACM international conference on Automated Software Engineering*, ASE '10, pages 427–436, New York, NY, USA, 2010. ACM.

An Overview of Digital Preservation Considerations for Production of “Preservable” e-Records: An Indian e-Government Case Study

Dinesh Katre

Centre for Development of Advanced Computing (C-DAC)
NSG IT Park, Aundh,
Pune 411007, India
91-20-25503386

dinesh@cdac.in

ABSTRACT

In the Indian context, when the e-government records are received for archival purpose, it is observed that very often they are produced without proper compliances for long term digital preservation. This paper presents a case study of e-district Mission Mode Project which offers diverse citizen services and produces the e-records such as birth certificates, domicile certificates, marriage certificates, caste certificates, etc in very large volumes. Such born digital e-government records have to be retained and preserved for technological and legal reasons. The Centre of Excellence for Digital Preservation established at C-DAC, Pune, India has carried out the study of e-record production process in the e-district and the need analysis for its digital preservation. The digital preservation best practices are identified, which have to be incorporated in the production process of e-records, so that the final e-records are produced in “preservable” form with full compliance as per the requirements of OAIS.

Insurance, National Citizen Database, Passport, Immigration, Visa and Foreigners Registration & Tracking, Pension, e-Office, Agriculture, Commercial Taxes, e-District, Employment Exchange, Land Records, Municipalities, Police, Road Transport, Treasuries. Citizen Service Centres, e-Biz, e-Courts, e-Procurement, etc.

The forthcoming Electronic Service Delivery Bill which is awaiting to be passed by the Indian parliament will make it mandatory for all government organizations and departments to offer the citizen services through electronic media within next 5 years. Enlarging volumes of e-records, e-documents and digital information are anticipated to be produced through these initiatives by the Indian government.

1.2 Legal framework

The Indian laws which clearly spell out the legal obligation of government organizations to preserve the electronic records are briefly introduced in this section.

1.2.1 Information Technology Act 2008

As per the IT Act, conditions for retention of electronic records are specified as - “accessibility so as to be usable for a subsequent reference; retention in the format in which it was originally generated, to represent accurately the information originally generated, with the details, which will facilitate the identification of the origin, destination, date and time of dispatch or receipt of such electronic record” [9].

1.2.2 Public Records Act 1993

As per the Public Records Act, every record creating agency of the central government, any ministry, department or office of the Government must provide proper arrangement, maintenance and preservation of public records [18].

1.2.3 Right To Information Act 2005

As per the Right To Information (RTI) Act, every public authority is obliged to maintain all its records duly catalogued and indexed and to ensure that all records that are appropriate to be computerized are, within a reasonable time, computerized and connected through a network all over the country on different systems so that access to such records is facilitated [20].

Apart from these, there are several other laws in the Indian constitution such as Copyright Act, Banker's Book Evidence Act, Indian Evidence Act (medico legal requirements) which also

Categories and Subject Descriptors

E.3 [Data Encryption] Public key cryptosystems

H.3.2 [Information Storage]: File organization, Record classification

I.7 [Document and Text Processing]: Document management, Document preparation, Format and notation, Markup Languages, Standards

J.1 [Administrative Data Processing]: Government

General Terms

Documentation, Design, Standardization, Theory, Legal Aspects.

Keywords

e-Government, Digital Preservation, Electronic Records, Fixed Digital Object, Significant Properties, Preservation Description Information (PDI), Submission Information Package (SIP), Open archival Information System (OAIS)

1. INTRODUCTION

1.1 Growth of e-records in India

The Indian government is spending more than 10 billion dollars on e-governance through its National e-Government Action Plan (NeGP) [2]. It has already launched 27 Mission Mode Projects which includes central, state and integrated MMPs such as Banking, Central Excise & Customs, Income Tax (IT),

emphasize the need to preserve the electronic records for various reasons.

1.3 India's National Digital Preservation Programme

The author of this paper was entrusted with the responsibility to prepare the National Study Report on Digital Preservation Requirements of India [16], as the first step towards formulating the Indian National Digital Preservation Programme of the Ministry of Information and Communications Technology, Government of India. The report included the recommendations of 30 experts from diverse domains across India. As per the study report, the Indian digital preservation scenario is observed as under-

- It is necessary to first establish what an e-record is in principle and how it can be recognized in electronic environment for preservation purpose [1, 3].
- The e-records are threatened by the continuing changes and obsolescence of computer hardware, software, file formats, storage media; and also the other dangers like data corruption, physical damage and disasters.
- There is lack of awareness about the need to preserve the e-records and the legal implications of failing to do so.
- There is absence of procedures and infrastructure for preserving the e-records.
- e-Government systems are being developed without incorporating the digital preservation consideration so that the e-records produced are preservable and comply with the minimum requirements of Open Archival Information System (OAIS).
- The present Departmental Record Officers (DROs), Record Keepers and Archivists working with the record producing agencies in India do not have the technical skills and knowledge of digital preservation [15, 17].

1.3.1 Centre of Excellence for Digital Preservation

Therefore, as per the recommendations given in the National Study Report on Digital Preservation Requirements of India, the Ministry of Information and Communications Technology, Government of India has funded the proposal of C-DAC Pune to establish the Centre of Excellence for Digital Preservation. This project aims at developing the standards, best practices, tools and systems for the preservation of electronic records. More information is available at <http://www.ndpp.in/>. The author of this paper is the chief investigator of this project.

2. RELATED WORK

Though there is limited guidance available on long term digital preservation of e-government records, we briefly discuss the most notable international projects related to this topic in this section.

The Canadian research project "International Research on Permanent Authentic Records in Electronic Systems (InterPARES) [8] offers the principles and guidance for the record creators and preservers both, so as to ensure the preservability of e-records when they are produced.

The following principles given by InterPARES are applied in our casestudy -

- The record creation process must be integrated with the recordkeeping rules with specific business processes [3].
- Digital objects must have a stable content and a fixed documentary form to be considered records and to be capable of being preserved over time.
- Preservation considerations should be embedded in all activities involved in record creation and maintenance if a creator wishes to maintain and preserve accurate and authentic records beyond its operational business needs.

National Archives Records Administration (NARA), USA provides record management guidance on digitally signed documents [19]. The following observation of NARA is particularly relevant to our case study –

- Since litigation will typically occur after the expiration of a public key certificate, it is important to take steps to ensure that pertinent records remain available after the certificate has expired. It is equally important that they be complete and understandable without the need for technical interpretation, to the extent possible.

Minnesota State Archives offers a broad strategy on E-records Management [4]. The National Archives of UK also provides the e-Government Policy Framework for Electronic Records Management.

However, the technical details and guidance provided by InterPARES and NARA were particularly helpful to us in understanding various aspects of e-records preservation. ISO/TR 15489 on Information and Documentation - Records Management is also very helpful in understanding the characteristics of records.

3. SCOPE

During our research on digital preservation of e-government records so far, we have come across following distinct categories of e-records –

▪ E-records with fixed information content

A process which culminates into a final certificate or an official document with fixed information content of long term importance. The final e-record is to be retained and used as it is, without requiring any further processing or alteration.

▪ Incrementally evolving e-records

A process in which new information is added into the e-record over a period e.g. banking transactions or change in the property ownership in land records. In such e-records the historical information of past transactions continues to be importance for preservation.

In this paper, we have focused on the digital preservation considerations for "final e-records with fixed information content" like birth certificate or domicile certificate issued to Indian citizens through the e-district Mission Mode Project (MMP).

4. CASE STUDY

We have chosen e-district Mission Mode Project (MMP) as a case study to build the pilot digital repository of e-records. The e-districts are offering following type of services to Indian citizens [21]-

- Creation and distribution of certificates for income, domicile, caste, Birth, Death etc.

- Arms Licenses, Driving Licenses, etc.
- Public Distribution System (PDS): Issue of Ration Card, etc.
- Social Welfare Schemes: Disbursement of old-age pensions, family pensions, widow pensions, etc.
- Marriage Registration, Land Records, etc.

Many services offered through e-district are producing large volumes of certificates which are authorized with digital signature. The certificates like birth certificate, marriage registration certificate, domicile certificate, caste certificate produced through electronic means need to be preserved as per the applicable retention rules and legal requirements. In this paper, we have focused on the digital preservation considerations related to certificates (birth, caste, marriage, domicile, etc) produced by e-districts.

After seeking due permissions, our team visited multiple e-districts, studied the system architecture and workflow, collected the sample database and certificates.

5. NEED ANALYSIS

The e-government systems should be designed to incorporate the following digital preservation considerations so as to produce the preservable e-records.

5.1 Need of e-record objectification

We observed that the e-district maintains a database comprising of various information elements and images pertaining to millions of certificates issued to various citizens. In one of the e-districts, the size of the database file was close to 3 TB, which is inflating everyday with the addition of new certificates issued to the citizens. The final certificate is dynamically rendered in the browser as per the layout specifications. The final certificate is not given an object form with fixed information content.

Figure 1. A database with data pertaining to millions of certificates

As per our assessment, the current approach poses following digital preservation challenges.

The digital information pertaining to certificates stored in the database is a result of the business logic which involves workflow, programme instructions, data structures, dependencies between values, formulas applied for calculated values and functions in force. Therefore any change in the business logic, representation logic and rendering logic can change the content of the certificate in an undesirable manner. Typically, the “current”

and transactional information should be maintained in the database. The final or “non-current” certificates should be given a fixed object form for long term preservation.

Refer figure 2 to understand the vulnerability to undesirable changes in e-records when they are under the influence of business logic.

Therefore, after the e-record is finalized, it is necessary to delink it from the business logic and fix it in the form of a self contained digital object for the purpose of preservation.

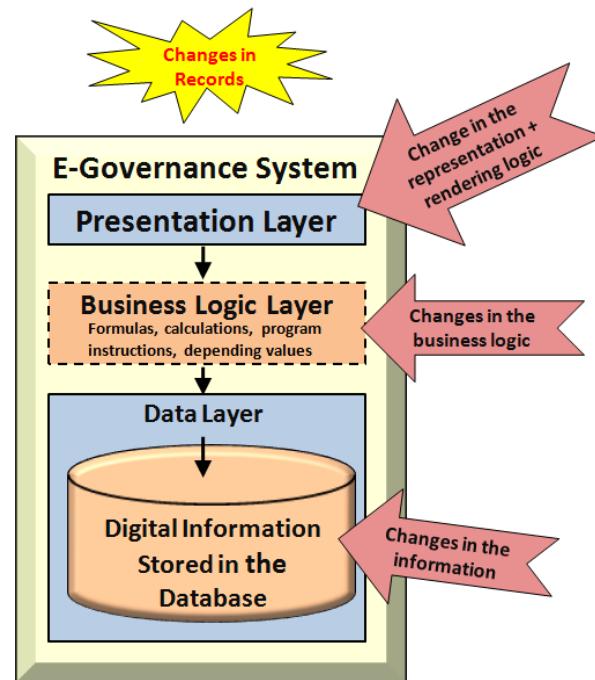


Figure 2. Vulnerability to undesirable changes in e-records

5.2 Need to digitally sign the entire certificate

Lets briefly understand the workflow of the e-district. The Citizen Service Centres (CSC) established in small towns and villages are connected to the e-district. The citizens are able to submit their application for certificate along with necessary documents and proofs with the CSC. The CSC operators digitize the applications and upload it for verification and issuing of certificate. The district authority verifies the documents and then grants the certificate (depending on the type of request) authorized by affixing the digital signature to “selected information values” (such as date of birth, name of person, etc) in the database. The digital signature is then stored in the database. The CSC is notified when the certificate is authorized. The approved information content is rendered in the browser as shown in figure 3 and then printed on standard stationary (paper) for issuing it to the citizen as shown in figure 4. It is a hybrid approach, in which the key contents of the certificate are born digital and digitally signed but the final certificate issued to the citizen is printed on paper.

In this process, many significant properties of the certificate such as layout, border, emblem, watermarked image, the authorization of state government regarding legal acceptance of digitally signed certificate are getting added only through the printed stationary. These significant properties are not part of the digitally signed information content of the certificate stored in the database.

The affixing of digital signature to selected information values in the database ensures its integrity but it does not certify or authorize the final certificate as shown in figure 4.

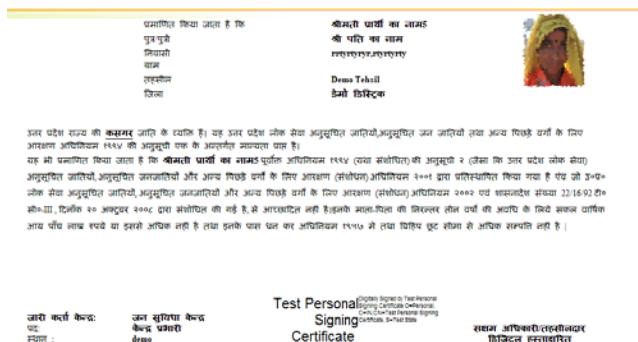


Figure 3. Rendering of a demo certificate with digital signature in browser



Figure 4. A digitally signed certificate printed on stationary paper

Although, our intention is not to comment on the process of authorization but from the digital preservation perspective, the difference between what remains in the database (refer figure 1) and what is issued as the final certificate (refer figure 3) is notable. Ideally the certificate issued to the citizen and the certificate retained for preservation must be exactly the same in terms of its logical and conceptual representations. To further substantiate this point, as per Duranti et al the form of transmission of a record is the physical and intellectual form that the record has when it is received; and the authenticity is best

ensured by guaranteeing that a record maintains the same form through transmission, both across space and through time [3].

5.3 Need of significant properties

The significant properties are those characteristics [technical, intellectual, and aesthetic] agreed by archive or by the collection manager to be the most important features to preserve over time [5]. In case of the certificates as shown in figure 4, the significant properties such as layout, border, emblem of the state government, font style for logo and color scheme are added only through the printed stationary. The dynamic on-screen rendering of certificate is dependent on browser and display settings. It may render differently on different computers. Therefore, the minimum essential significant properties of the certificates must be purposefully designed and embedded in its digital rendering and given a fixed form for long term preservation. The significant properties are helpful to the curators in asserting or demonstrating the continued authenticity of objects over time, or across transformation processes [7].

5.4 Need of file naming policy

It is observed that the file names generated by the e-government systems follow some type of incremental numbering system but such filenames are not adequate to be consistent, meaningful, unique and parseable. For example, the files pertaining to birth certificates, domicile certificates, marriage certificates, etc can be categorized by pre-fixing a standard code or short forms such as BC, DC and MC in the file name. It will be so helpful in categorizing the certificates based on file names.

5.5 Need of Preservation Description Information (PDI) along with certificates

It is observed that most of the e-government projects are focused on offering the citizen services but no consideration is given to how the e-records produced by the e-government systems will be preserved for future. It is possible to capture some parts of the Preservation Description Information (PDI) through e-government system itself while producing the final e-record or the certificate. The final digital object must accompany the PDI with minimum essential metadata for it to be acceptable as a valid Submission Information Package (SIP) for the Open Archival Information System (OAIS) [12]. If we consider the huge volumes of e-records it is not practically possible to generate the PDI in a post facto mode at the time of archival and therefore, we suggest that it should be automatically captured when the e-record is produced.

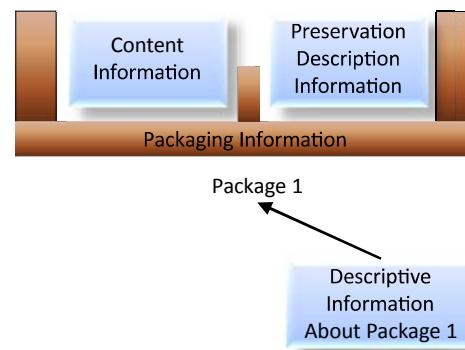


Figure 5. Need of Preservation Description Information (PDI) as per OAIS

If the certificates produced through e-district were to be discovered, read and understood in future then we will need to know the following –

- What is the identifier of certificate?
- To whom was it issued?
- When, where and who had produced it?
- What was the context in which it was produced?
- What was the basis on which the certificate was issued?
- Which software was used for producing the certificate?
- In which file format was it stored?
- How to know that the certificate available in the archive is the authentic one?
- What can be given as the proof or evidence of its authenticity?
- How to know if the certificate has not been modified?
- Does it require to be converted in the latest file format to be able to render it and read it?
- Who is authorized to access and read the certificate?

The answers to these questions are normally found in the Preservation Description Information (PDI).

Due to unavailability of the standard vocabulary and metadata schema, the present e-government systems are not able to produce the PDI along with the final e-record.

In this context, while exploring the ways in which PDI could be described for e-records, we came across the application of DSpace for cataloging of court case records which were described using Dublin Core Metadata Elements as shown in figure 6.

One can notice that the names appearing in front of dc.contributor.author are the names of judges who passed the final judgment on the court case.

The names appearing in front of dc.contributor.editor and dc.contributor.illustrator are the names of Petitioner and Respondent in the particular court case.

Full metadata record		
DC Field	Value	Language
dc.contributor.author	N.K.PATIL AND K.N.KESHAVANARAYANA	en_US
dc.contributor.editor	S B KRISHNAMURTHY S/O BOREGOWDA	en_US
dc.contributor.illustrator	K N NARASIMHAIAH S/O NARASANNA	en_US
dc.coverage.spatial	Bangalore	en_US
dc.date.accessioned	2009-10-23T08:14:33Z	-
dc.date.available	2009-10-23T08:14:33Z	-
dc.date.created	2009-10-05	-
dc.date.issued	2009-10-23T08:14:33Z	-
dc.identifier.isbn	2001	en_US
dc.identifier.issn	2006	en_US
dc.identifier.sici	8	en_US
dc.identifier.ismn	MFA	en_US
dc.identifier.uri	http://hdl.handle.net/123456789/209929	-
dc.description.abstract	MFA 9233/2005	en_US
dc.title	MFA 2001/2006	en_US
Appears in Collections:	Miscl. First Appeal - MFA	
Files in This Item:		
File	Description	Size Format
MFA2001-06-05-10-2009.pdf	346.45 KB	Adobe PDF View/Open
Show simple item record		

Figure 6. DCMES applied through DSpace to describe a court case record

It is obvious that the Dublin Core Metadata Elements are more suitable for describing the resources like books and not suitable for court cases or certificates or e-government records. It is also very misleading, as the judges are mapped as the authors, petitioner is mapped as the editor, and the respondent is mapped as the illustrator. Also, the court cases do not have ISBN.

Therefore, a suitable metadata schema with appropriate vocabulary (which represents the local understanding) is needed for the description of certificates and e-government records in Indian context.

The requirements identified so far are part of the packaging process involved in the making of a Trustworthy Digital Object (TDO) [6].

6. BEST PRACTICES AND GUIDELINES

Based on the study of workflow and characteristics of e-government records (certificates) produced through e-district, the Centre of Excellence for Digital Preservation has identified following best practices and guidelines for production of preservable e-records.

6.1 The final certificate as a fixed digital object

As per the findings of Canadian InterPARES 2 (International Research on Permanent Authentic Records in Electronic Systems) project, the preservation considerations should be embedded in all activities involved in record creation and maintenance if a creator wishes to maintain and preserve accurate and authentic records beyond its operational business needs. ISO/TR 15489-2 for Information Documentation - Records Management Guidelines also specifies the need to capture the e-record with fixed representation of actions [13]. Therefore, the final contents (information + images + significant properties) of the certificate produced by e-district should be given a composite and fixed object form.

Selection criteria for objectification of e-record

The e-records should be produced in the form of a fixed digital object on the basis of following criteria-

- The e-record is meant to be used as a certificate or a final statement proof
- The legal obligations and implications of failing to reproduce such e-record in its original and authentic form in future
- The value of information contained in the e-record
- The e-record forming a basis or dependency for other transactions
- The historical significance of the e-record
- The retention rules pertaining to such e-records
- The record keeping and preservation policy of the record producing organization

Typically the e-records like birth certificate, domicile certificate, marriage registration certificate, death certificate, senior citizen certificate, insurance policy, ration card, passport, income tax return, mark sheet, service record or documents such as MoU, contract, agreement, parliamentary bills / acts, court case judgments along with proceedings, user manuals which need to be retained for various reasons (like legal, value of information,

historical importance) can be considered to be produced in the form of a digital object with fixed information content.

6.1.1 The criteria for not giving a fixed object form to e-records-

The following type of e-records need not be given an object form based on following criteria-

- The e-record has temporary significance
- There are no legal obligations or implications for not maintaining such e-record beyond its purpose of use
- As per the retention rules such e-record is not required for more than 5 years (in that case it can be maintained in the database)

6.2 The PDF for Archival (PDFA) format specification for final certificate

As per our study, some e-government systems are producing the proprietary Adobe PDF output which is not recommended for preservation. Therefore, the final e-records like certificates should be objectified in the form of PDF for Archival format specified as under-

ISO 19005 PDFA-1a is recommended for archival of “born digital documents” [10].

ISO 19005 PDFA-1b is recommended for archival of “reformatted digital documents” (for example composite PDF comprising of TIFF images).

PDFA-2a [11] can also be used but PDFA-1a and PDFA-1b is adequate in the present context.

6.3 Conceptual representation of certificate

An e-record in the database is nothing else but digital information distributed in various tables of the database. It forms the logical representation of the given e-record. The conceptual representation of e-record covers the rendering attributes and visual appearance which are essential for human sensorial understanding of the e-record. Most e-government systems are designed to store the logical representation of e-records. Such systems do not address the requirements of the conceptual representation of e-records which is necessary to be captured while producing the final digital object in the PDF/A format.

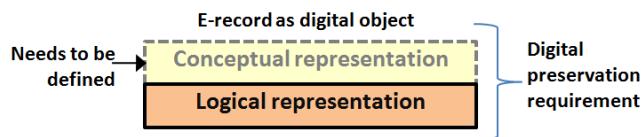


Figure 7. Need to specify the conceptual representation of e-record

We need to specify the significant properties of e-record so that its conceptual representation or the rendering aspects get properly addressed.

6.3.1 Significant properties for certificates

Following type of significant properties should be embedded in the final e-record at the time of objectification.

- Proper page layout (page size, orientation, margins)

- Tables with specifically defined columns, rows and cell spacing
- Emblem / logo of the organization with proper color specification / color code
- Header and footer information
- Font specifications, style settings for titles and the textual information
- Bar code
- QR code
- Images with specific DPI, dimensions and format
- Watermarked image
- Fixed location coordinates for images
- Fixed location coordinates for digital signature

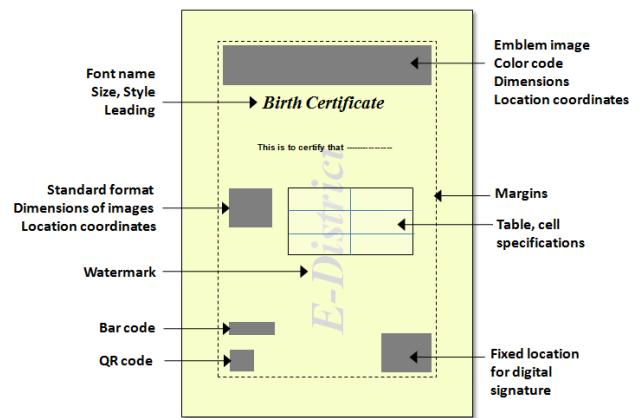


Figure 8. Significant properties of a certificate

6.3.2 Why significant properties are important for preservation of e-records?

The significant properties are extremely helpful in fulfilling the requirements of usage, authentication, preservation and several legal obligations which are enlisted below-

- Legal obligation as stated in IT ACT 2000 (b) – “the electronic record is retained in the format in which it was originally generated, sent or received or in a format which can be demonstrated to represent accurately the information originally generated” [9].
- ISO 14721:2003 OAIS [12] specifies and refers this requirement as Information Properties needed for preservation.
- Meaningful understanding and usage of information
- Verification of originality and authenticity of e-record
- Renderability of contents exactly as original in future, even if the present document format or software becomes obsolete
- Reconstructability of the digital object by using its elements

6.4 Consistent and logical file naming policy

The record producer (e-district) can select at least 3 to 4 relevant file name elements as per the examples given in this section for defining the logical and consistent file naming policy.

Appropriate abbreviations / short forms can be used along with separators and incremental serial numbers. We must avoid using the controlled characters and empty spaces in filenames. The filename / length / character sets should be compatible across operating systems / file systems. Examples of file name elements are given below-

- Type of certificate
- Service code
- Reference number / accession number
- Place
- Date of creation
- Name of creator / organization
- Title of content
- Department number
- Name of organization
- Records series

6.5 Affix the digital signature to final e-record in PDFA format

- After completing all information processing the final e-record is produced in the form of PDFA document and then it should be digitally signed by the competent authority for authorization and non-repudiation.
- The PDFA document could be printed on the standard stationary paper for issuing to the citizen.
- The PDFA document is then submitted for archival and preservation, which has the required significant properties.

6.6 Capture the Preservation Description Information (PDI) of final e-record during its production process

The Centre of Excellence for Digital Preservation has defined a comprehensive metadata schema titled as “E-governance Standard for Preservation Information Documentation of E-records (E-Gov SPIDeR) based on the types of e-records produced in the Indian context.

We have studied the existing metadata schemas like Dublin Core, MODS, METS and PREMIS. The designers of these metadata schemas have considered wide range of objects and it reflects the state-of-the-art and maturity of archiving practices in the developed countries. As per our assessment, the existing metadata schemas are too exhaustive and not perfectly fitting in the context of Indian e-government records.

We needed something smaller, simpler and yet comprehensive which could capture the minimum essential preservation information at the time of record production itself. Therefore, we have defined our own metadata schema for the description of e-records which reflects our local understanding and requirements. It is a hybrid metadata schema which includes our own contributions in addition to the selected metadata elements from the established schemas.

The major sections of the e-Gov SPIDeR metadata schema are briefly explained here as it is not possible to reproduce the entire schema due to space limitation.

▪ Cataloging Information

The cataloging metadata for e-records retains some of the Dublin Core metadata elements with new additions like RecordIdentifier, RecordType, MainCategory, SubCategory, NameID, OfficeType, Validity and RetentionDuration. The Paris Principles for cataloguing [14] are adopted for defining the common cataloguing parameters for electronic records [13].

▪ Enclosure Information

The final e-record (e.g. the certificates issued by e-district) is generated on the basis of various documents, proofs and correspondence which are enclosed with it. The enclosure information is needed for establishing the context in which the e-record was produced. The list of enclosures can be included in the PDI if applicable. The accuracy of the final e-record can be verified and validated on the basis of the enclosed documents.

▪ Provenance Information

It includes the address of Citizen Service Centre (CSC) that received the application for certificate, the office address of e-district which issued the final certificate and the device IDs of the servers where the request was processed and final certificate was issued.

▪ Representation Information

It includes the names and version information of software, operating system, compiler, API Library, application, tools, web browser, database, etc which was used for creating the final e-record and the software necessary for reading it.

▪ Fixity Information

It includes the checksums of the final e-record (certificate) and its enclosures.

▪ Digital Signature Information

Digital signature metadata portion is adopted from PREMIS.

▪ Access Rights Information

The access rights metadata portion is adopted from METS.

It is ensured that the E-Gov SPIDeR metadata schema can be mapped with the established metadata schemas like Dublin Core.

As per our study, major portion of the metadata can be captured at the time of record production itself as the required descriptive information is either getting generated through the process or it is available in the database.

If this information is not captured during the record production then it is likely to remain scattered in e-government systems, and eventually it may be lost forever. Also, in the post facto mode it is difficult to gather the descriptive metadata for ingest and ensure its accuracy and authenticity.

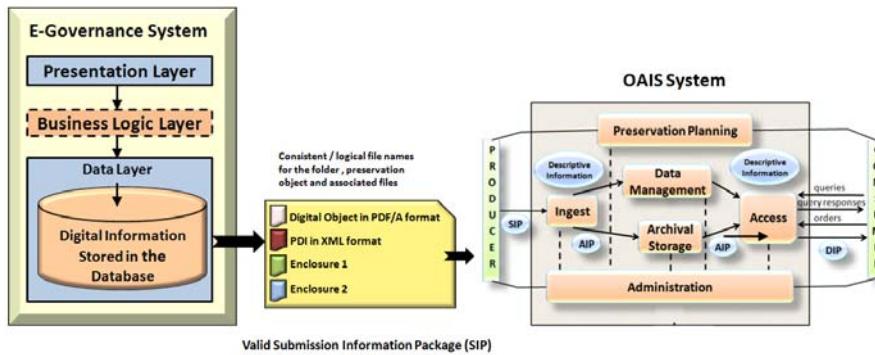


Figure 9. Final e-record produced with basic digital preservation consideration

7. CONCLUSION

In case of e-Government Records, it is necessary to incorporate the basic digital preservation considerations throughout the e-record production. It is important to ensure that the e-government systems are designed and developed in such a way that the final e-records produced by them are “preservable” enough and comply with the requirements of the OAIS standard [12].

8. ACKNOWLEDGMENTS

The encouragement and support received from Department of Electronics and Information Technology, Government of India and C-DAC Pune is acknowledged with gratitude and thankfulness.

9. REFERENCES

- [1] Acland G. I., 1996. Electronic Records: The View From Beyond OZ, Australian Society of Archivists Conference, Alice Springs, (May 1996).
- [2] All Govt. services to go online by 2014, 2010. Deccan Herald, Bangalore, Saturday, (Jul. 2010) 8.
http://www.mit.gov.in/sites/upload_files/dit/files/ApexMeeting_23072010_0.pdf
- [3] Duranti L. and MacNeil H., 1996. The Protection of the Integrity of Electronic Records: An Overview of the UBC-MAS Research Project, *Archivaria* 42 (Fall 1996): 46-67.
- [4] Electronic Records Management Guidelines, Minnesota State Archives
http://www.mnhs.org/preserve/records/electronicrecords/erm_s.html
- [5] Giaretta D., 2011. Advanced Digital Preservation, 1st Edition, Published by Springer Verlag, Berlin Heidelberg (June 2011), ISBN-10: 3642168086.
- [6] Gladney H. M., 2010. Preserving Digital Information, Published by Springer Verlag, Berlin Heidelberg (2010), ISBN 978-3-642-07239-0.
- [7] Grace S., Knight G. and Montague L. 2009. Final Report on Investigating the Significant Properties of Electronic Content over Time (InSPECT), The National Archives of UK, (December 2009).
<http://www.significantproperties.org.uk/inspect-finalreport.pdf>
- [8] InterPARES 2, 2008. International Research on Permanent Authentic Records, A Framework of Principles for the Development of Policies, Strategies and Standards for the Long-term Preservation of Digital Records (March 2008)
[http://www.interpares.org/public_documents/ip2\(pub\)/policy_framework_document.pdf](http://www.interpares.org/public_documents/ip2(pub)/policy_framework_document.pdf)
- [9] Information Technology Act, 2008.
- [10] ISO 19005-1:2005 PDF/A-1
- [11] ISO 19005-2:2011 PDF/A-2
- [12] ISO 14721:2003 Open Archival Information Systems (OAIS)
- [13] ISO/TR 15489-1 and 2 Information and Documentation - Records Management
- [14] International Conference on Cataloguing Principles (Paris : 1961). Report. – London : International Federation of Library Associations, 1963, p. 91-96.
- [15] Katre D. S., 2009. Ecosystem for digital preservation in Indian context: A proposal for sustainable and iterative lifecycle model. In Proceedings of Indo-US Workshop on International Trends in Digital Preservation, (March 2009), Pune, India, 137–141.
<http://ndpp.in/download/Indo-US-DP-Proceedings-C-DAC-2009.pdf>
- [16] Katre D. S., 2010. National Study Report on Digital Preservation, Requirements of India. Volume I: Recommendations for National Digital Preservation Programme, Published by C-DAC, India, (2010).
- [17] Katre D. S., 2011. Digital preservation: converging and diverging factors of libraries, archives and museums - An Indian perspective, *IFLA Journal*, Vol. 37, no. 3, (October 2011), Sage Publications, London, UK, 195-203.
- [18] Public Records Act, 1993.
- [19] Records Management Guidance For PKI Digital Signature Authenticated and Secured Transaction Records by Federal Public Key Infrastructure Steering Committee Legal/Policy Working Group, National Archives and Records Administration, (March 2005).
<http://www.archives.gov/records-mgmt/policy/pki.html#4-5>
- [20] Right To Information Act, 2005.

Developing Research Data Management Capability: the View from a National Support Service

Sarah Jones

University of Glasgow
HATII, 11 University Gardens
Glasgow, G12 8QJ
+44 141 330 3549
sarah.jones@glasgow.ac.uk

Graham Pryor

University of Edinburgh
DCC, Appleton Tower
Edinburgh, EH8 9LE
+44 131 650 9985
graham.pryor@ed.ac.uk

Angus Whyte

University of Edinburgh
DCC, Appleton Tower
Edinburgh, EH8 9LE
+44 131 650 9986
a.whyte@ed.ac.uk

ABSTRACT

An increasing number of UK Higher Education Institutions (HEIs) are developing Research Data Management (RDM) support services. Their action reflects a changing technical, social and political environment, guided by principles set out in the Research Councils UK (RCUK) Common Principles on Data Policy. These reiterate expectations that publicly-funded research should be openly accessible, requiring that research data are effectively managed. The Engineering and Physical Sciences Research Council (EPSRC) policy framework is particularly significant, as it sets a timeframe for institutions to develop and implement a roadmap for research data management.

The UK Digital Curation Centre (DCC) is responding to such changes by supporting universities to develop their capacity and capability for research data management. This paper describes an ‘institutional engagement’ programme, identifying our approach, and providing examples of work undertaken with UK universities to develop and implement RDM services. We are working with twenty-one HEIs over an eighteen month period, across a range of institution types, with a balance in research strengths and geographic spread. The support provided varies based on needs, but may include advocacy and awareness raising, defining user requirements, policy development, piloting tools and training. Through this programme we will develop a service model for institutional support and a transferable RDM toolkit.

Categories and Subject Descriptors

E.0 [Data General].

General Terms

Management, Design, Security, Human Factors, Legal Aspects.

Keywords

Research Data Management, data sharing, university, higher education, infrastructure, research data policy, Data Management

Plan, training, Digital Curation Centre, JISC.

1. INTRODUCTION

The desire among UK Higher Education Institutions (HEIs) to develop Research Data Management (RDM) roadmaps is driven by a range of factors. Developments in research data policy are a key influence, as are social and political demands for transparency. Controversies sparked by prominent Freedom of Information requests for research data have had a detrimental effect on institutional reputations and brought the risks of poor data management into sharp focus. Concurrently ‘data driven’ technologies have reshaped the research process and demonstrated benefits of scale and impact in a growing number of disciplines.

Reflecting the broader changes noted above, the JISC-funded Digital Curation Centre (DCC) supports the UK higher education community to manage, curate and preserve digital material. Most recently, DCC effort has been focused on managing research data. We distinguish RDM from preservation by the former’s emphasis on verifiable and replicable processes to support research data use from its planning, through its creation and active use, to its point of handover to a repository or archive. These include preservation actions to ensure fitness for access, use and reuse, as described for example in the DCC Curation Lifecycle Model [1]

Research data management represents new demands for HEIs in terms of technical and organisational infrastructure, the provision of specialist data curation skills and long term planning for sustainable services. We are currently working with twenty-one HEIs through our institutional engagement programme to increase their RDM capability in these areas whilst developing a support model that can be redeployed with other UK universities charged with facing what are commonly seen as additional technological and policy challenges.

There are two key outputs from the DCC institutional engagement programme: 1) a model for supporting HEIs to develop their RDM capabilities, i.e. their ability to articulate and achieve RDM objectives; and 2) a transferable RDM toolkit. The support model is outlined in section 3. It involves applying tools to help initiate processes of change in each institution, diagnosing current practice, and implementing redesigned services. The RDM toolkit describes potential HEI services, examples of which are given in section 4. These include exemplars of DMP Online, an online data management planning tool customised for HEIs by using ‘institutional templates’. Each of the HEIs the DCC is supporting

has agreed to share their experience and to allow others to reuse outputs from our engagement with them.

2. DATA POLICY BACKGROUND

An increasing number of HEIs are developing policies and implementation plans for research data management. These are often guided by funder requirements and codes of good research practice. The Engineering and Physical Sciences Research Council's (EPSRC) policy framework for research data, which was released in May 2011, places the onus on institutions to address research data management. It sets out clear timescales for implementation: research organisations should develop a roadmap to align their policies and processes with EPSRC expectations by 1st May 2012, and be fully compliant with them by 1st May 2015 [2].

Most research councils have released similar policies promoting the effective management and open sharing of research data. The RCUK Common Principles on Data Policy [3] highlight the importance of policies and plans – both at institutional and project-specific level. Importantly, they also confirm that it is appropriate to use public funds to support management and sharing of publicly-funded research data, enabling the development of support infrastructure.

A trend for institutional research data management policies is evident. A number of policies emerged in 2011-2012 and many more are in draft form awaiting approval, as listed by the DCC [4]. These policies frame the institutional governance needed to develop associated infrastructure and embed good practice. The policies tend to be accompanied by guidelines for implementation or more detailed local policies and processes. Data Management Plans (DMPs) written for specific projects or as group guidelines play an important role in this framework. Six of the seven UK research councils expect researchers to submit DMPs in grant proposals, while the seventh (EPSRC) advocates the importance of plans but does not require their submission.

3. A MODEL FOR SUPPORTING HEIs

Our model for supporting HEIs is being refined by implementing it through the engagement process. We first outline the scope of the two main tools we are applying: DAF (Data Asset Framework) and CARDIO (Collaborative Assessment of Research Data Infrastructure and Objectives), which both originate from digital preservation research and development projects. We then describe three business process change stages that we aim to contribute to in each HEI: initiating change; diagnosing data practices; and (re)designing services. We identify the role of DAF, CARDIO and other tools relevant to each stage.

We describe how the engagements fit within the support model, as shown in Figure 1. This comprises two other ongoing activities; evaluation of each engagement, and comparison across them. These result in forthcoming outputs; firstly reports describing and evaluating each engagement, and our comparisons of these across institutions. The latter will document our refined model, based on improved understanding of how best to deploy the DAF and CARDIO tools to develop institutional capabilities, and factors enabling and inhibiting this. The second main output planned is a transferable RDM ‘toolkit’ of service descriptions, exemplars and good practice guidance that other institutions can

deploy. This includes exemplars of support for Data Management Planning, where localised services have been developed.

3.1 Tools for Engagement

Each ‘institutional engagement’ aims to build the institution’s capability by working with them to articulate the need for change, and scope requirements for redesigned services. We envisage institutional services will combine technology with ‘soft’ infrastructure including training, guidelines, and policies to support these [5], i.e. the changes needed may be at least as much of an organisational nature as a technical one.

We deploy a range of tools and approaches developed through recent collaborative projects. Two DCC tools have supported the initial work: -

Collaborative Assessment of Research Data Infrastructures and Objectives (CARDIO) aims to help establish consensus on RDM capabilities and gaps in current provision. Institutional preparedness is self-assessed using a capability model adapted for RDM from the ‘three legged stool’ model of Cornell University Library’s digital preservation programme [6]. Users rate existing provision in three areas - organisation, technology & resources - and come together to agree the ratings and prioritise action. The tool can be used online, in person or a combination of these.

The *Data Asset Framework (DAF)* is a survey and interview-based methodology to investigate research groups’ data holdings and how these are managed. Questionnaires and interviews generally cover the range of activities involved in the curation lifecycle to identify issues and gaps. DAF has been piloted in a number of contexts through case studies [7].

3.2 Developing Institutional Infrastructure

Our assumption is that formally structuring and coordinating data management can benefit research. Nevertheless we take the introduction of effective RDM as a rubric for bringing change to a range of highly diverse activities. Sociotechnical research demonstrates the complexity of developing infrastructure in the context of diverse and changing requirements, and the necessity for both short and long-term views to be included in planning this development [e.g. 8]. We see RDM infrastructure development as a process of change that requires input from at least three perspectives; research practice, management, and information systems development. These perspectives may come from an institutions’ Library, IT and Research Support functions, as well as from researchers themselves.

Institutional RDM service development can be viewed as an iterative cycle similar to business process redesign. Ideally long-term planning should be encompassed in a process of learning and continuing improvement. Our initial focus is on early stages of process redesign, which we adapt from Kettinger et al’s framework [9]; initiating change, diagnosing data practices, and redesigning services¹.

¹ Our current process emphasizes steps two to three in the six stages identified by Kettinger et al [9]; envision, initiate, diagnose, redesign, reconstruct and evaluate.

We characterize the three stages below, indicating relevant DCC methods and tools alongside other examples. The support described is aimed at scoping the redesign of support services and roles. As RDM matures as an element of UK institutions' service provision we anticipate that support for implementing new systems and evaluating services may require further modeling tools to relate these to, for example, enterprise IT architecture (e.g. [10]).

3.2.1 Initiating change

This stage is led by a champion authorised by senior management to form a steering group to scope a project and enlist academic engagement. In several cases, a member of the senior management group responsible for research chairs this, e.g. the Deputy/Pro Vice Chancellor for Research. A steering group would typically consider research strategies, service priorities and technology opportunities, and identify stakeholders, issues and domains to investigate. Having planned and secured the necessary resources for a project, including the human resource to plan and implement change, their initial work is likely to focus on raising stakeholder awareness and obtaining "buy-in".

Engaging senior researchers will be vital given the differences between research and an institution's administrative processes, research practices being more diverse and fluid. Senior management support is also needed to ensure that strategy is aligned with feasible action, given the competing demands for resources and a constrained funding environment. Policy development may be needed to communicate institutional priorities and define responsibilities. Benchmarking to identify capability gaps and analysis of risks and benefits may help the case for change and identify the main goals and success factors.

Methods/tools: DAF, CARDIO, KRDS/I2S2 Benefits Analysis Tool [11].

3.2.2 Diagnosing data practices

The next stage involves profiling data management and sharing norms, roles and values, aiming to identify the main issues encountered by researchers and other service users or stakeholders. Typically a project manager or steering group member with operational responsibility will undertake this work in a series of short studies, involving selected research groups and providers of any relevant existing services such as backup storage or library support.

The aim here is to appreciate enough about current RDM practices, their shaping by disciplinary factors, and usage of available sources of support, to identify the appetite for change, how needs are framed, and the likely barriers to aligning them with strategy and regulatory requirements. The diagnostic stage may therefore include assessment of the awareness of relevant policies, and chart the lifecycle of typical data assets and associated research objects (software, protocols, logs, etc).

Methods/tools: The DAF approach aims to support this form of enquiry into typical data lifecycles, stakeholders involved, and their concerns and priorities. CARDIO complements this by identifying service providers' assessment of current provision. Other tools and methods relevant here include Data Curation Profiles [12], and Stakeholder Profiles [13]. Benefits frameworks may help identify priorities, e.g. the KRDS/I2S2 Benefits Analysis Tool (*ibid*). Where there is substantial existing support for data archiving and a need for more detailed analysis of workflows, Research Activity Information Development (RAID) diagrams provide a modeling tool to support this [14].

3.2.3 Redesigning research data services

This stage involves the project manager and any operational group working with stakeholders to describe new service options, and their feasibility and desirability. The tools relevant here will

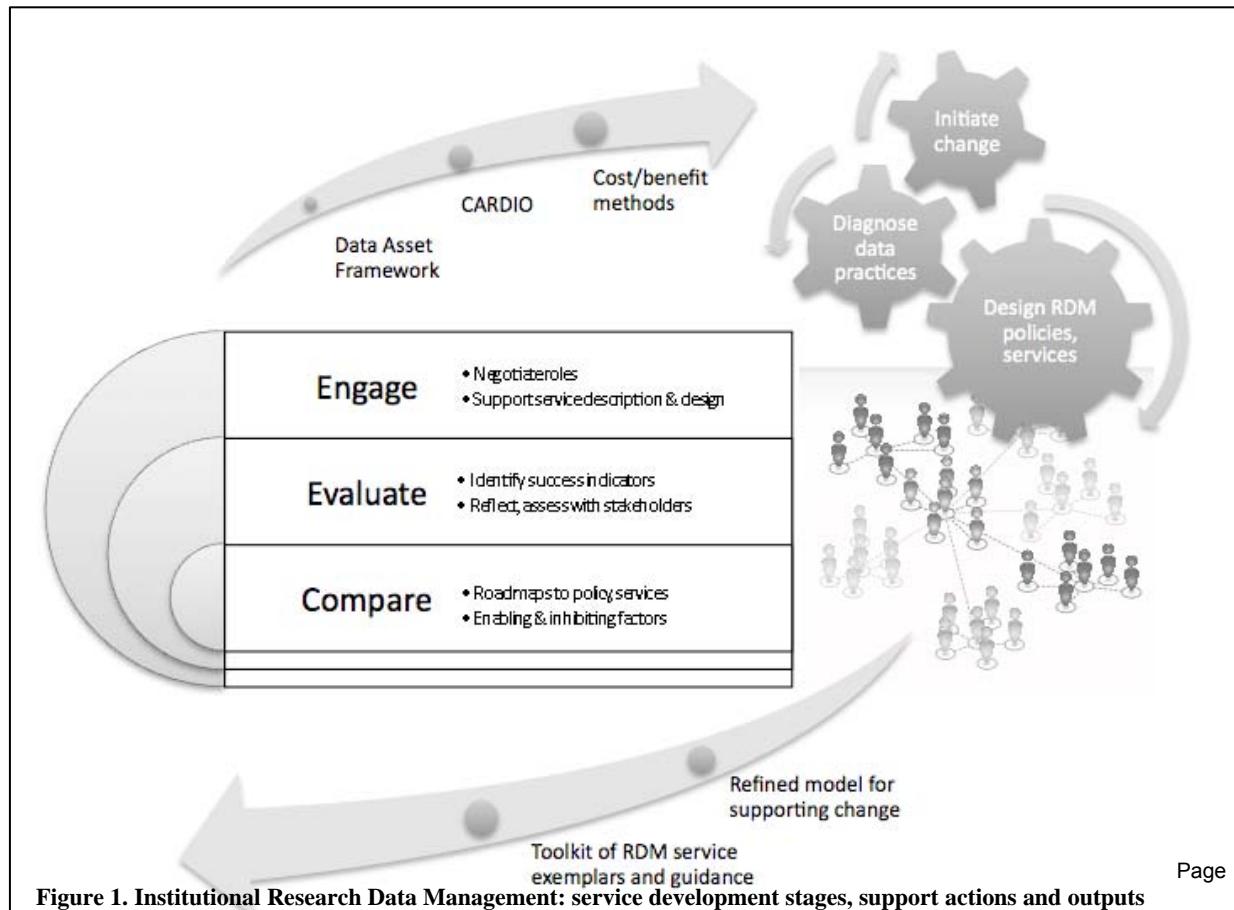


Figure 1. Institutional Research Data Management: service development stages, support actions and outputs

vary with the kind of service proposed. Frameworks for Institution-wide policies and guidelines may take the form of exemplars drawn from other institutions. This is likely to involve senior academics, along with Research Office and or Library (e.g. academic liaison) colleagues, in assessing options according to needs scoped using CARDIO or DAF. Similar stakeholders will be involved in defining training needs, and here too the CARDIO and DAF tools should have highlighted the policy areas and RDM concepts that training needs to raise awareness of.

Further tools will be needed when the options for change involve developing new information systems, or include requirements to interoperate with existing systems e.g. institutional repositories, or research information management systems. This will involve business analysts from Library or IT systems areas. Tool support may be needed to articulate new process concepts. This may for example use scenarios to present narrative ‘user stories’ and use cases. Workflows diagrams (e.g. the RAID method outlined above) and prototypes may help the intended users and stakeholders to compare ‘as-is’ and ‘to-be’ processes, whether on cost/benefit or other criteria e.g. research ethics or strategic objectives.

Support for researchers to develop a data management plan (DMP) when applying for research funding is likely to be one component of service provision. DCC provides a tool, *DMP Online* [15] that contributes here by providing templates and guidance to encourage good practice. *DMP Online* originates from a checklist to help researchers meet funders’ grant application requirements [16]. The tool can be adapted to individual institutions and our experiences in doing so through the engagement programme provides further lessons in the drivers and barriers to implementing RDM services.

Methods/ tools: RAID diagrams, DMP Online tool, Stakeholder profiles, Soft Systems Methodology [17].

3.3 Evaluating and Comparing Engagements

Each institution provides a mini-case study of factors decisive in shaping institutional research data management; and each offers opportunities to refine the DCC tools and the use cases for delivering these either as generic web-based applications, or as bespoke offerings used with substantial DCC mediation.

Action research methods emphasise learning methodically from involvement in problem solving, and are appropriate given that the DCC programme is funded as capacity building rather than research *per se*. Soft Systems Methodology (SSM) is an action research method consistent with our assumptions about the need for ‘soft’ infrastructure. While we do not claim to follow the approach rigorously, the important aspects for our purpose are:

- Identifying and engaging with stakeholders who are articulating the need for change
- Appreciating how they frame relevant issues and contexts
- Providing opportunities to articulate feasible and desirable service improvements

To these ends, the authors and colleagues participate in institutions’ steering groups, hold workshops with stakeholders to discuss findings, and provide training in good data management practice. To support cross-institutional comparisons the

programme holds internal workshops to reflect critically on factors enabling and inhibiting success, across the institutions and our interventions to support them. This also benefits from participation in external workshops held by the JISC Managing Research Data programme, which is funding institutions to conduct similar organisational change and service development projects [18].

Key questions guide our evaluations, whose overall aim is to refine the support model with our stakeholders input, and compare individual engagements. Our key questions are:

1. What stakeholders become engaged in RDM service development, and what new roles are adopted?
2. What are common priorities for RDM services, and enablers and barriers to developing these?
3. How much intermediation is needed to use DAF and CARDIO and how may these best be used in combination?
4. What are our client’s and stakeholders’ success indicators, and how do they assess our contribution?

The rest of the paper addresses the first two questions, and we conclude on the scope of the RDM service toolkit.

4. PROGRAMME PARTICIPANTS

The engagement programme was promoted to institutions via the DCC’s data management roadshows [19]. These are regional events whose main aim is to bring stakeholders together to address institutional RDM issues. The roadshows have encouraged interest, and most of the engagements were initiated through them. For example in a recent roadshow a local institution’s, Head of Internet Services, Library Academic Services Manager, and Head of Research Development came together to develop a strategy. They subsequently approached the DCC for assistance and we are defining a programme of support.

The level of interest in the programme has allowed us to establish a balanced portfolio. The twenty-one HEIs currently taking part are spread geographically across the UK and represent a range of university types. Three participants are ancient universities, formed in the 15th and 16th centuries. Another six participants are civic institutions with origins dating from the 19th and early 20th centuries. Eight were formed in the 1960s, while the remainder are former polytechnics that became universities post 1992.

We have sought participation of universities with a variety of research portfolios and strategies. Six of our participants are members of the Russell Group, which “represents the 20 major research-intensive universities of the UK” [20]. Several others are known for particular research strengths and bring these to the portfolio. The more modern institutions focus primarily on teaching but have ambitions to develop their research profile.

The EPSRC policy has been a key impetus for institutions to form working groups with the intention of developing RDM strategies. As we expected these involve a range of services, typically the library, IT and research office. The lead partner in the majority of our engagements is the library. Indeed every engagement has some representation from the library; in cases where they are not leading, library-based staff often undertake the majority of the work. The research office is leading in seven of the cases and is

involved in most of the others. Institutional IT Services are only leading in two of the engagements. Furthermore, IT involvement is lacking in a few other engagements, raising questions about how effectively technical change can be embedded.

5. INITIAL FINDINGS

If the institutions the DCC is supporting can be seen as representative, then UK universities are in the early stages of addressing research data management. Most are scoping requirements and benchmarking current practice to plan future work. We are aware of few institutions where components of an RDM infrastructure or support services are already in place. Most are early in the process of developing services.

The following sections highlight some key areas of activity.

5.1 Research Data Management Policies

Many of the participating institutions have responded to the trend to develop research data management policy. Requests for support have ranged from feedback on drafts to developing policies on their behalf. The DCC has provided a policy briefing [21] in support of this activity, which outlines requirements and summarises different approaches that universities have taken. A number of institutions have looked to the University of Edinburgh's seminal policy developed in 2010-11 [22] and used that as a base from which to adapt.

The DCC capitalized on broader interest in this area by inviting participants to join a JISC Managing Research Data (MRD) policy workshop was held in March 2012, which provided an opportunity to share practice and learn from others. Key discussion points were the degree of specificity needed and the optimum timing of an RDM policy. Questions were raised about the level of detail required of an institutional policy leading to suggestions for more detailed implementation guides and tailored departmental policies. Fears were also expressed about approving RDM policies before the associated infrastructure was in place to make compliance problematic.

Only three participants have RDM policies that pre-date DCC involvement. In these cases the emphasis of our work is on policy implementation. Pilot studies are being run with researchers at one institution to see how easily they can write a data management plan and deposit data for preservation and sharing, as outlined in the policy.

5.1.1 Example A: Policy development

One participating university created a Research Services librarian post in 2011. This post aimed to support researchers, in line with the institution's ambition to be a leading modern university for research. The person recruited was tasked to lead the University's RDM initiatives. As in other institutions, the EPSRC expectations were a driving force.

The initial task in this university was to develop an RDM policy. Existing policies were reviewed in November 2011. A first draft was largely based on the University of Edinburgh policy, with additions to define further responsibilities and agree periodic review dates for data retention. Feedback from a small focus group was positive; researchers sought clarification on the scope and wanted practical guidance for implementation. There was also a desire that the policy should be supportive rather than strongly

enforced. With researcher support, the policy was put forward to the Research & Knowledge Exchange Committee and approved.

The policy development and approval at this institution, a relatively new university, took four months. In part this is due to existing examples that could be repurposed, drastically reducing the effort needed in composition. The process of approval was also far simpler than in older universities, which tend to have various committee levels that need to be passed.

5.2 Roadmaps and Strategy Development

The EPSRC's policy [2] places a number of expectations on institutions. They must ensure awareness of the policy and regulatory framework for RDM, identify internal data holdings, publish metadata about these, and provide infrastructure to preserve them. The policy calls for long-term commitment to preservation; institutions are required to keep selected datasets accessible for at least ten years from the end of any embargo period, or from the date of the last third party access request. Institutions must also define responsibilities for curation activity across the DCC Curation Lifecycle.

The EPSRC expects 'roadmaps' to plan RDM infrastructure and services and ensure compliance with their expectations. This has provided the context for the DCC to help participating institutions scope a response. Institutions should define the content and format of the roadmap and initially self-assess their compliance. However the EPSRC has made clear that future funding may depend on inspection and compliance. This has provided an impetus for our work with RDM steering groups.

This work has drawn primarily on the DAF (Data Assessment Framework) and CARDIO (Collaborative Assessment of Research Data Infrastructure and Objectives) tools. Typically steering groups have preferred CARDIO where their institutions have a range of relevant services in place. Other steering groups have preferred to conduct DAF surveys or interviews to gather evidence of current awareness and needs. These have been carried out through pilot groups, identified with varying degrees of DCC support. The pilots provide evidence for developing roadmaps and policy, and a model for the steering groups to apply with further groups across their institution.

DAF questionnaires have been tailored to suit institutional circumstances; in some cases they have been used online and as the basis for structured interviews; in others as topic guides for semi-structured interviews. In some cases steering group members have undertaken the interviews themselves, with DCC advising on questions and format, and in others they have shadowed the DCC staff doing interviews and, having gained familiarity with the topics and structure, taken a more active role in later interviews.

We have provided workshops at the beginning of these pilot studies, often combining RDM training and awareness raising sessions with introductions to the DAF approach. We also use workshops towards the end to communicate and consolidate results. CARDIO has been used for both purposes; some institutions have opted to use it to benchmark service provision before further investigation, others to take stock of the results of the investigation.

5.2.1 Example B: Roadmap development

One institution's steering group is led by its Research Office and Records Management staff. Our role was to propose a roadmap format and gather evidence for initial self-assessments. Initially this involved helping to define pilot groups in two faculties, and then carrying out DAF interviews with researchers at various levels of seniority, from doctoral students to research group leaders. These profiled current practice, gauged demands for change, and informed a CARDIO gap analysis against EPSRC expectations. The interviews with researchers and support staff across faculties also highlighted gaps between expectations that other funders place directly on researchers themselves, and the support available to help researchers meet these.

It proved useful to organise the roadmap under the headings of training, policy development, service development and policy implementation, and finance. This helped separate tasks that could be accomplished in the short term, from others requiring additional roles and resources. Short-term requirements included basic RDM training to be embedded in postgraduate training. Longer term requirements included systems for cataloguing active research data at faculty and/or research group level, guidelines and processes for appraising and selecting material of long-term value, and identifying the appropriate place to deposit/preserve it or ensure appropriate disposal.

5.3 Data Management Planning

The DCC web-based tool to assist in this process, DMP Online, has three main functions: to help create and maintain different versions of DMPs; to provide useful guidance on data management issues and how to meet research funders' requirements; and to export useful plans in a variety of formats. The tool draws upon the DCC's analysis of funders' data requirements to help project teams create two iterations of a data management plan: an 'application' stage plan and a 'funded' plan.

Several of the institutions in the programme have asked for a tailored version of DMP Online. This enables universities to add customised guidance, such as links to relevant webpages and contact details for support staff. A new feature in v3.0 of the tool is the ability to provide suggested answers: universities can compose text for inclusion in cases where generic provision is in place, such as central storage and backup. Customised versions of DMP Online incorporate the institution's logo and can be branded to apply relevant design and URLs so they are seen as an institutional service.

5.3.1 Example C: Customising DMP Online

At one participating university DCC support is part of the institution's IT Transformation project, which is addressing various aspects of research data management, including storage and tools. Some preliminary work on data management planning was undertaken by a JISC-funded project in a research centre in the university. This provided the catalyst for a customisation of DMP Online.

A preliminary meeting was held early in 2012 to discuss requirements with the project manager. The process of customization was explained and a schedule agreed. An implementation team at the institution has documented requirements and produced an institutional template based on the

elements of the DCC Checklist, which they wish to include together with details of local support. The DCC has input this information to create the template in the tool, and supported ongoing user testing. Training materials are being developed to suit this institution's context and a launch is planned for 2012.

5.4 Managing Research Data Storage

Managing storage is a primary concern for researchers, and as such is high on the list of priorities for universities. Activity is typically focused on providing sufficient quantities of research data storage. Tools to enable data sharing with external collaborators and version control are also sought. Analogies are often made with Dropbox when describing requirements [23].

Significant developments in this area are being made in the wider community. The DataFlow project at the University of Oxford [24] is one of a range of RDM applications resourced by the Higher Education Funding Council for England (HEFCE), as potential cloud-based services for universities. DataFlow is a two-stage data management infrastructure intended to make it easy for researchers to work with, annotate, share, publish, and permanently store their research data. There are two components: DataStage, a secure, local file management system with private shared and collaborative directories, and DataBank, a scalable data repository designed for institutional deployment. Several of the universities DCC is supporting have flagged an interest in piloting DataFlow.

5.4.1 Example D: Data storage strategy

At one institution a Vice Principal convened two working groups to progress their RDM initiatives: one on research data management and one on research data storage. The research data storage working group identified requirements for a cross-platform file store, accessibility for external collaborators, and provision for backup and synchronisation. Requirements were also identified for services to deliver data archiving and federated data storage.

A business case was made and resources released to purchase infrastructure and develop support services. The DCC has assisted the working group to develop a list of existing and proposed services. Pilot studies are planned to test the different ways forward. The expectation is that existing provision will be extended to allocate a nominal 0.5TB per researcher, with provision co-ordinated at local level.

5.5 Guidance and Training

DAF and CARDIO studies have uncovered a discrepancy between existing support provision and awareness of this. In many cases collating details of existing services and improving their presentation presents a 'quick win'. This was done on the JISC-funded Incremental project at the Universities of Cambridge and Glasgow, and provides a useful model for redeployment [25]. Short, simple guidance tends to be called for, as data management can seem overwhelming if presented in a technical way.

Training of some kind features in over 25% of the engagements. There are two key areas of interest: disciplinary courses for PhD students and professional training to re-skill research support staff. Our emphasis is on extensively reusing existing resources. The DCC's DC101 course [26] and Data Intelligence 4 Librarians

[27] by the 3TU consortium in the Netherlands are both targeted at research support staff. The JISC RDMTrain projects produced disciplinary courses [28] and the UK Data Archive has also produced training materials for researchers [29].

The DCC provides bespoke training courses by adapting relevant resources to specific institutional needs. Requirements currently being addressed include provision for one institution's academic liaison librarians to introduce RDM to researchers; and in another institution providing content for PhD training in Health and Life Sciences.

5.5.1 Example E: Training development

One participating university has also been running a JISC MRD infrastructure project. In collaboration with that team, we have supported a number of training initiatives. Training is run via the Doctoral Training Centre, with the hope that by catching young researchers early, you can instill good data management habits before they start to make bad ones. We trained the most recent cohort at the beginning of the academic year and they have supported one another since. The training gave a grounding in research data management and used data management plans as a vehicle to put the principles into practice. The PhD students trialed a number of DMP templates to see which was most appropriate to develop a plan to guide their work.

6. CONCLUSIONS AND NEXT STEPS

The DCC model for supporting institutions to build RDM capacity is working well. In all cases a plan of action has been developed with a steering group and is in the process of being delivered. For the majority of institutions this has involved diagnosing current practice to define requirements, as most were unaware of their current position at the outset. A few institutions are nearing completion of the DCC engagement, having indicated they feel equipped to continue development themselves.

In many cases, the initial stages have taken some time to build momentum, as the process of reaching consensus and initiating change can be daunting. However, progress has been far quicker in some institutions than others. This appears to be due to a range of factors. A few institutions have committed resource to research data management and funded a position to spearhead activity and build momentum. There also seems to be a quicker process of change in smaller, more modern institutions. This could in part be due to their structure: fewer levels of hierarchy make it easier to raise ideas and elicit approval. Cultural factors may also be at play: smaller scales can make it easier to engage the research community and there appears to be a greater willingness amongst researchers to work with central services.

Some approaches have worked particularly well, such as our focus on engaging early career researchers in training, in the expectation that they will filter change upwards as they permanently adopt good data practice as part of their routine research process. With policies defined and the benefits explained, institutions are also beginning to grapple with the creation of business plans designed to ensure that the necessary technical and human infrastructures are sustainable. We have found the principal concerns across participating institutions' steering groups to be similar. The main indicators of success for them are the formulation of roadmaps to address compliance requirements, which are common to all UK universities (e.g.

[30]), and 'quick wins' in terms of responses to researchers' demand for clear guidance and easily managed storage provision.

Despite the DCC emphasis on providing generic web-based solutions, engagement demands flexibility and adaptation to local contexts. Most of the usage of DAF and CARDIO has been with our mediation, and this has enabled us to identify needs to improve the flexibility and integration of online tools to support this. Evaluation to prioritise specific improvements in tools and methods (e.g. workshop formats) is ongoing, comprising telephone interviews with participating stakeholders, and usability assessments of the online elements of support provision.

Our next steps include cross-institutional surveys on the needs for support in policy compliance, and the degree to which involvement in our programme has supported this. A number of important differences have already emerged through cross-site comparison. Requirements for support vary: some researchers create vast quantities of complex data and require improved storage management to make analysis scalable. For others, the challenges are more in the heterogeneity of data form. Attitudes to data sharing set others apart: those working with human subjects require tightly controlled access, whereas other groups have adopted a culture of data sharing and demand easier external collaboration. Requirements can be diverse across and within disciplines, so a flexible approach is needed.

From a data curation perspective one should not exaggerate the differences. Despite them we find that similar issues apply in supporting data management: policy development and planning, training and guidance, data management planning, managing storage for active research data, data evaluation/appraisal, gathering and publishing metadata, identifying relevant external repositories, choosing repository platforms, systems integration, managing data access and citation, and making the case for long-term sustainability. Many of these issues overlap with preservation, and in supporting active research data management we continue to draw lessons from the preservation community.

On a national level, these are still relatively early days in the change process. Continued support will be needed over the coming years as pilot projects transition into embedded services. For the DCC, the formal conclusion of each sixty-day engagement is not the end of our collaboration. Continuity in support is vital to a community that is fluid by nature and notorious for the speed with which initiatives decay when the driving force is removed before the achievement of critical mass.

The outputs of the DCC engagement programme are adding to a growing body of exemplars that can be repurposed. Parallel work in the JISC MRD programmes, data centres and RDM initiatives in a number of UK universities are similarly providing RDM service exemplars and outputs that can be repurposed. The key for institutions is to draw relevant aspects from these examples, which suit their research culture and environment. The DCC engagement programme aims to provide an adaptive framework for doing this. We hope to refine and share this framework beyond the borders of the engagement programme as a model for other HEIs to improve their research data management practice.

7. ACKNOWLEDGMENTS

We acknowledge the support of the Higher Education Funding Council for England (HEFCE), and the JISC via the Universities Modernisation Fund. Our thanks are also extended to the various

universities that have participated. Final reports from the programme will name the institutions concerned and incorporate their comments. We also thank three anonymous reviewers.

This paper represents work undertaken by a DCC team comprising Brian Aitken, Alex Ball, Michael Day, Martin Donnelly, Monica Duke, Marieke Guy, Patrick McCann, Andrew McHugh, Kerry Miller, Jonathan Rans, and the authors.

8. REFERENCES

- [1] Higgins, S. 2008. The DCC Curation Lifecycle Model. *International Journal of Digital Curation* 3 (1)
- [2] EPSRC. 2011. Policy Framework on Research Data: Expectations. <http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx>
- [3] Research Councils UK. 2011. Common Principles on Data Policy. <http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx>
- [4] Digital Curation Centre. 2012. Institutional Data Policies. <http://www.dcc.ac.uk/resources/policy-and-legal/institutional-data-policies/uk-institutional-data-policies>
- [5] Ward, C., Freiman, L. Jones, S. Molloy, L. and Snow, K. 2011. Making Sense: Talking Data Management with Researchers. *International Journal of Digital Curation* 6 (2) (October 7). doi:10.2218/ijdc.v6i2.202. <http://www.ijdc.net/index.php/ijdc/article/view/197>
- [6] Kenney, A. and McGovern, N. 2005. The Three-Legged Stool: Institutional Response to Digital Preservation. II Convocatoria del Coloquio de marzo. Cuba. http://www.library.cornell.edu/iris/dpo/docs/Cuba-arknym_final.ppt
- [7] Jones, S., A. Ball, and Ç Ekmekcioglu. 2008. The Data Audit Framework: a First Step in the Data Management Challenge. *International Journal of Digital Curation* 3 (2): 112–120.
- [8] Whyte, A. 2012. Emerging infrastructure and services for research data management and curation in the UK and Europe. In *Managing Research Data*, G. Pryor, Ed. Facet Publishing, London. 173-204.
- [9] Kettinger, W. J., Teng, J.T.C and Guha, S. 1997. Business Process Change: A Study of Methodologies, Techniques, and Tools. *MIS Quarterly* 21 (1) (March 1): 55–80. doi:10.2307/2449742.
- [10] Becker, C., Antunes, G., Barateiro, J. and Vieira, R. A Capability Model for Digital Preservation. In Proc. iPRES 2011, 2011.
- [11] Beagrie, N. 2011. KRDS/I2S2 Benefits Analysis Tool. <http://beagrie.com/krds-i2s2.php>
- [12] Witt, M., Carlson, J. Brandt, D. S and Cragin, M. H 2009. Constructing Data Curation Profiles. *International Journal of Digital Curation* 4 (3).
- [13] Michener, W. K., Allard, S. Budden, A. Cook, R.B., Douglass, K. Frame, M. Kelling, S., Koskela, R., Tenopir, C., and David A. Vieglais, D.A.. Participatory Design of DataONE—Enabling Cyberinfrastructure for the Biological and Environmental Sciences. *Ecological Informatics* (0). doi:10.1016/j.ecoinf.2011.08.007.
- [14] Darlington, M., Ball, A. Howard, T, Culley, S. and McMahon, C. 2011. RAID Associative Tool Requirements Specification (February 23). <http://opus.bath.ac.uk/22811/>
- [15] Donnelly, M., Jones, S and Pattenden-Fail, J.W. 2010. DMP Online: The Digital Curation Centre's Web-based Tool for Creating, Maintaining and Exporting Data Management Plans. *International Journal of Digital Curation* 5 (1) (June 22): 187–193. doi:10.2218/ijdc.v5i1.152.
- [16] Digital Curation Centre. 2011. Checklist for a Data Management Plan http://www.dcc.ac.uk/webfm_send/431
- [17] Checkland, P., and Poulter, J. 2010. Soft Systems Methodology. In *Systems Approaches to Managing Change: A Practical Guide*, Ed. Reynolds, M. and Holwell, S. London: Springer London. 191–242. <http://www.springerlink.com.ezproxy.webfeat.lib.ed.ac.uk/content/p35875774m64g2l2/>.
- [18] Joint Information Systems Council. 2011. Managing Research Data Programme 2011-13. http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.aspx
- [19] Digital Curation Centre. 2011. Data Management Roadshows <http://www.dcc.ac.uk/events/data-management-roadshows>
- [20] Russell Group. 2012. About Us. <http://www.russellgroup.ac.uk/about-russell-group/>
- [21] Digital Curation Centre. 2011. Research data policy briefing <http://www.dcc.ac.uk/resources/policy-and-legal>
- [22] University of Edinburgh. 2011. Research Data Management Policy. <http://www.ed.ac.uk/schools-departments/information-services/about/policies-and-regulations/research-data-policy>
- [23] Cope, Jez. 2012. MRD Hack Days: File backup, sync and versioning, or “The Academic Dropbox” <http://blogs.bath.ac.uk/research360/2012/05/mrd-hack-days-file-backup-sync-and-versioning-or-the-academic-dropbox>
- [24] DataFlow. 2012. What we’re doing <http://www.dataflow.ox.ac.uk/index.php/datastage/users/researhers>
- [25] University of Cambridge. 2010. Incremental Scoping Study Report and Implementation Plan. <http://www.lib.cam.ac.uk/preservation/incremental/news.html>
- [26] Digital Curation 101 <http://www.dcc.ac.uk/training/dc-101>
- [27] Data Intelligence 4 Librarians <http://dataintelligence.3tu.nl/en/home>
- [28] JISC RDMTrain outputs <http://www.dcc.ac.uk/training/train-trainer/disciplinary-rdm-training/disciplinary-rdm-training>
- [29] Managing and Sharing data course <http://www.dataarchive.ac.uk/create-manage/training-resources>
- [30] Research Councils UK. 2009. RCUK Policy and Code of Conduct on the Governance of Good Research Conduct. <http://www.rcuk.ac.uk/review/grc/default.htm>

Advancing Data Integrity in a Digital Preservation Archive Ex Libris and the Church of Jesus Christ of Latter-day Saints

Nir Sherwinter (nir.sherwinter@exlibrisgroup.com) and Gary T. Wright (wrightgt@ldschurch.org)

1. INTRODUCTION TO THE CHURCH OF JESUS CHRIST OF LATTER-DAY SAINTS

The Church of Jesus Christ of Latter-day Saints is a worldwide Christian church with more than 14.4 million members and 28,784 congregations. With headquarters in Salt Lake City, Utah (USA), the Church operates three universities, a business college, 138 temples, and thousands of seminaries and institutes of religion around the world that enroll more than 700,000 students in religious training.

The Church has a scriptural mandate to keep records of its proceedings and preserve them for future generations. Accordingly, the Church has been creating and keeping records since 1830, when it was organized. A Church Historian's Office was formed in the 1840s, and later it was renamed the Church History Department.

Today, the Church History Department has ultimate responsibility for preserving records of enduring value that originate from the Church's ecclesiastical leaders, Church members, various Church departments, the Church's educational institutions, and its affiliations.

With such a broad range of record sources within the Church, the array of digital record types requiring preservation is also extensive. However, the vast majority of storage capacity in the Church's digital preservation archive is allocated to audiovisual records.

Over the last two decades, the Church has developed state-of-the-art digital audiovisual capabilities to support its vast, worldwide communications needs. One such need is broadcasting semiannual sessions of General Conference, which are broadcast in high definition video via satellite to more than 7,400 Church buildings in 102 countries and are simultaneously translated into 32 languages. Ultimately, surround sound digital audio tracks for more than 90 languages are created to augment the digital video taping of each meeting—making the Church the world's largest broadcaster of languages.

Another communications need is producing weekly broadcasts of *Music and the Spoken Word*—the world's longest continuous network broadcast (now in its 84th year). Each broadcast features an inspirational message and music performed by the Mormon Tabernacle Choir. The broadcast is aired live by certain radio and television stations and is distributed to approximately 2000 other stations for delayed broadcast.

The Church's Publishing Services Department, which supports all these broadcasts, generates multiple petabytes of production audiovisual data annually. In just ten years, Publishing Services anticipates that it will have generated a cumulative archival capacity of more than 100 petabytes for a single copy.

2. INTRODUCTION TO THE EX LIBRIS GROUP

Ex Libris is a leading provider of library automation solutions, offering the only comprehensive product suite for the discovery, management, distribution, and preservation of digital materials. Dedicated to developing the market's most inventive and creative solutions, Ex Libris products serve the needs of academic, research, national, and other libraries, such as the Church History Library. With more than 460 employees worldwide, Ex Libris operates an extensive network of eleven wholly-owned subsidiaries and twelve distributors, many of which are exclusive. Ex Libris corporate headquarters are located in Jerusalem, Israel.

3. BUILDING THE CHURCH'S DIGITAL RECORDS PRESERVATION SYSTEM

In order to build and maintain a large digital archive, the Church History digital preservation team realized that it would be critical to minimize the total cost of ownership of archival storage.

An internal study was performed to compare the costs of acquisition, power, data center floor space, maintenance, and administration to archive hundreds of petabytes of digital records using disk arrays, optical disks, virtual tape libraries, and automated tape cartridges. The model also incorporated assumptions about increasing storage densities of these different storage technologies over time.

Calculating all costs over a ten year period, the study concluded that the total cost of ownership of automated tape cartridges would be 33.7% of the next closest storage technology (which was disk arrays). Consequently, the Church uses IBM 3500 Tape Libraries with LTO-5 and TS1140 tape drives for its digital preservation archive today.

Another requirement was scalability. Clearly, a multi-petabyte archive requires a system architecture that enables rapid scaling of automated ingest, archive storage capacity, access, and periodic validation of archive data integrity.

After several discussions with qualified, relevant people, concerns over the ability of open source repositories to adequately scale eliminated these potential solutions from consideration.

Ex Libris Rosetta was evaluated next. In order to determine if it would be able to scale to meet Church needs, a scalability proof of concept test was conducted.

The Rosetta evaluation involved joint scalability testing between Ex Libris and the Church History Department. Results of this testing have been published on the Ex Libris website (exlibrisgroup.com). The white paper is titled "The Ability to Preserve a Large Volume of Digital Assets—A Scaling Proof of Concept."

Results of the scalability test indicated that Rosetta would be able to meet Church History needs.

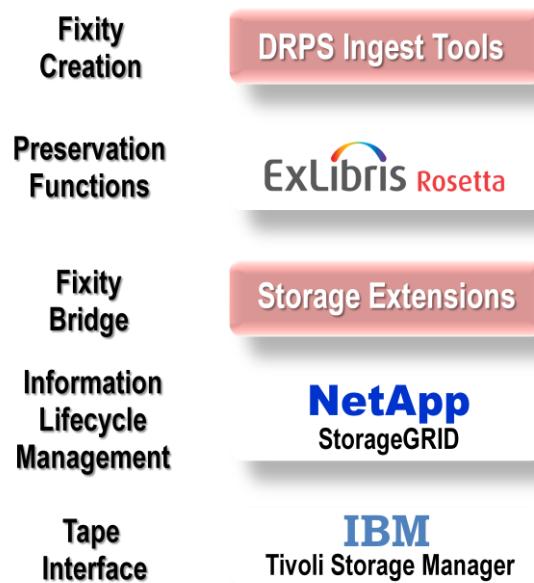
Next, the digital preservation team implemented the Church History Interim Preservation System (or CHIPS) using Rosetta for a more comprehensive test. CHIPS used only disk for storage. When the CHIPS proof of concept test was completed with successful results, the Church History Department decided to move forward with Rosetta as the foundation for its Digital Records Preservation System (DRPS—see Figure 1).

Rosetta provides configurable preservation workflows and advanced preservation planning functions, but only writes a single copy of an Archival Information Package [1] (AIP—the basic archival unit) to a storage device for permanent storage. An appropriate storage layer must be integrated with Rosetta in order to provide the full capabilities of a digital preservation archive, including AIP replication.

After investigating a host of potential storage layer solutions, the preservation team chose NetApp StorageGRID to provide the Information Lifecycle Management (ILM) capabilities that were desired. In particular, StorageGRID's data integrity, data resilience, and data replication capabilities were attractive.

In order to support ILM migration of AIPs from disk to tape, StorageGRID utilizes IBM Tivoli Storage Manager (TSM) as an interface to tape libraries.

DRPS also employs software extensions developed by preservation team members from Church Information and Communications Services (shown in the reddish boxes in Figure 1). These software extensions will be discussed later.



*Figure 1
Components of the Church History Department's
Digital Records Preservation System
(DRPS)*

4. DATA CORRUPTION IN A DIGITAL PRESERVATION TAPE ARCHIVE

A critical requirement of a digital preservation system is the ability to continuously ensure data integrity of its archive. This requirement differentiates a tape archive from other tape farms.

Modern IT equipment—including servers, storage, network switches and routers—incorporate advanced features to minimize data corruption. Nevertheless, undetected errors still occur for a variety of reasons. Whenever data files are written, read, stored, transmitted over a network, or processed, there is a small but real possibility that corruption will occur. Causes range from hardware and software failures to network transmission failures and interruptions. Bit flips (also called bit rot) within data stored on tape also cause data corruption.

Recently, data integrity of the entire DRPS tape archive was validated. This validation run encountered a 3.3×10^{-14} bit error rate.

Likewise, the USC Shoah Foundation Institute for Visual History and Education has observed a 2.3×10^{-14} bit error rate within its tape archive, which required the preservation team to flip back 1500 bits per 8 petabytes of archive capacity. [2]

These real life measurements—one taken from a large archive and the other from a relatively small archive—provide a credible estimation of the amount of data corruption that will occur in a digital preservation tape archive. Therefore, working solutions must be implemented to detect and correct these data errors.

5. DRPS SOLUTIONS TO DATA CORRUPTION

In order to continuously ensure data integrity of its tape archive, DRPS employs fixity information.

Fixity information is a checksum (i.e., an integrity value) calculated by a secure hash algorithm to ensure data integrity of an AIP file throughout preservation workflows and after the file has been written to the archive.

By comparing fixity values before and after files are written, transferred across a network, moved, or copied, DRPS can determine if data corruption has taken place during the workflow or while the AIP is stored in the archive. DRPS uses a variety of hash values, cyclic redundancy check values, and error-correcting codes for such fixity information.

In order to implement fixity information as early as possible in the preservation process, and thus minimize data errors, DRPS provides ingest tools developed by Church Information and Communications Services (ICS) that create SHA-1 fixity information for producer files *before* they are transferred to DRPS for ingest (see Figure 1).

Within Rosetta, SHA-1 fixity checks are performed three times—(i) when the deposit server receives a Submission Information Package (SIP) [1], (ii) during the SIP validation process, and (iii) when an AIP file is moved to permanent storage. Rosetta also provides the capability to perform fixity checks on files after they have been written to permanent storage, but the ILM features of StorageGRID do not utilize this capability. Therefore, StorageGRID must take over control of the fixity information once files have been ingested into the grid.

By collaborating with Ex Libris on this process, ICS and Ex Libris have been successful in making the fixity information hand off from Rosetta to StorageGRID.

This is accomplished with a web service developed by ICS that retrieves SHA-1 hash values generated independently by StorageGRID when the files are written to the StorageGRID gateway node. Ex Libris developed a Rosetta plug-in that calls this web service and compares the StorageGRID SHA-1 hash values with those in the Rosetta database, which are known to be correct.

Turning now to the storage layer of DRPS, StorageGRID is constructed around the concept of object storage. To ensure object data integrity, StorageGRID provides a layered and overlapping set of protection domains that guard against data corruption and alteration of files that are written to the grid.

The highest level domain utilizes the SHA-1 fixity information discussed above. A SHA-1 hash value is generated for each AIP (or object) that Rosetta writes to permanent storage (i.e., to StorageGRID). Also called the Object Hash, the SHA-1 hash value is self-contained and requires no external information for verification.

Each object contains a SHA-1 object hash of the StorageGRID formatted data that comprise the object. The object hash is generated when the object is created (i.e., when the gateway node writes it to the first storage node).

To assure data integrity, the object hash is verified every time the object is stored and accessed. Furthermore, a background verification process uses the SHA-1 object hash to verify that the object, while stored on disk, has neither become corrupted nor has been altered by tampering.

Underneath the SHA-1 object hash domain, StorageGRID also generates a Content Hash when the object is created. Since objects consist of AIP data plus StorageGRID metadata, the content hash provides additional protection for AIP files.

Because the content hash is not self-contained, it requires external information for verification, and therefore is checked only when the object is accessed.

Each StorageGRID object has a third and fourth domain of data protection applied, and two different types of protection are utilized.

First, a cyclic redundancy check (CRC) checksum is added that can be quickly computed to verify that the object has not been corrupted or accidentally altered. This CRC enables a verification process that minimizes resource use, but is not secure against deliberate alteration.

Second, a hash-based message authentication code (HMAC) message authentication digest is appended. This message digest can be verified using the HMAC key that is stored as part of the metadata managed by StorageGRID. Although the HMAC message digest takes more resources to implement than the CRC checksum described above, it is secure against all forms of tampering as long as the HMAC key is protected.

The CRC checksum is verified during every StorageGRID object operation—i.e., store, retrieve, transmit, receive, access, and background verification. But, as with the content hash, the HMAC message digest is only verified when the object is accessed.

Once a file has been correctly written to a StorageGRID storage node (i.e., its data integrity has been ensured through both SHA-1 object hash and CRC fixity checks), StorageGRID invokes the TSM Client running on the archive node server in order to write the file to tape.

As this happens, the SHA-1 (object hash) fixity information is not handed off to TSM. Rather, it is superseded with new fixity information composed of various cyclic redundancy check values and error-correcting codes that provide *TSM end-to-end logical block protection* when writing the file to tape.

Thus the DRPS fixity information chain of control is altered when StorageGRID invokes TSM; nevertheless, validation of the file's data integrity continues seamlessly until the file is written to tape.

The process begins when the TSM client appends a CRC value to file data that is to be sent to the TSM server during a client session. As part of this session, the TSM server performs a CRC operation on the data and compares its value with the value calculated by the client. Such CRC value checking continues until the file has been successfully sent over the network to the TSM server—with its data integrity validated.

Next, the TSM server calculates and appends a CRC value to each logical block of the file before transferring it to a tape drive for writing. Each appended CRC is called the “original data CRC” for that logical block.

When the tape drive receives a logical block, it computes its own CRC for the data and compares it to the original data CRC. If an error is detected, a check condition is generated, forcing a re-drive or a permanent error—effectively guaranteeing protection of the logical block during transfer.

In addition, as the logical block is loaded into the tape drive’s main data buffer, two other processes occur—

(1) Data received at the buffer is cycled back through an on-the-fly verifier that once again validates the original data CRC. Any introduced error will again force a re-drive or a permanent error.

(2) In parallel, a Reed-Solomon error-correcting code (ECC) is computed and appended to the data. Referred to as the “C1 code,” this ECC protects data integrity of the logical block as it goes through additional formatting steps—including the addition of an additional ECC, referred to as the “C2 code.”

As part of these formatting steps, the C1 code is checked every time data is read from the data buffer. Thus, protection of the original data CRC is essentially transformed to protection from the more powerful C1 code.

Finally, the data is read from the main buffer and is written to tape using a read-while-write process. During this process, the just written data is read back from tape and loaded into the main data buffer so the C1 code can be checked once again to verify the written data.

A successful read-while-write operation assures that no data corruption has occurred from the time the file’s logical block was transferred from the TSM client until it is written to tape. And using these ECCs and CRCs, the tape drive can validate logical blocks at full line speed as they are being written!

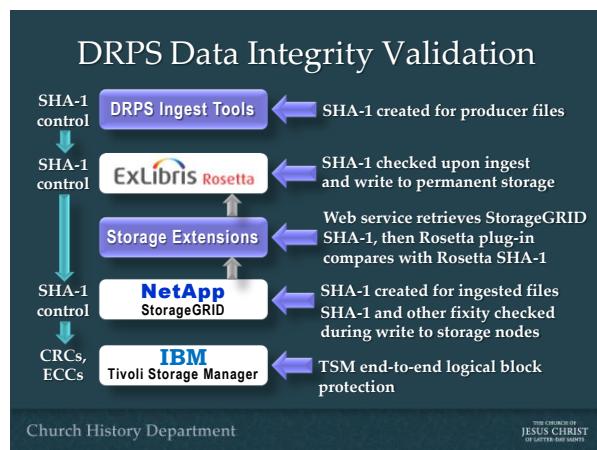
During a read operation (i.e., when Rosetta accesses an AIP), data is read from the tape and all three codes (C1, C2, and the original data CRC) are decoded and checked, and a read error is generated if any process indicates an error.

The original data CRC is then appended to the logical block when it is transferred to the TSM server so it can be independently verified by that server, thus completing the TSM end-to-end logical block protection cycle.

This advanced and highly efficient TSM end-to-end logical block protection is enabled with state-of-the-art functions available with IBM LTO-5 and TS1140 tape drives.

When the TSM server sends the data over the network to a TSM client, CRC checking is done once again to ensure integrity of the data as it is written to the StorageGRID storage node.

From there, StorageGRID fixity checking occurs, as explained previously for object access—including content hash and HMAC message digest checking—until the data is transferred to Rosetta for delivery to its requestor, thus completing the DRPS data integrity validation cycle.



*Figure 2
Summary of the DRPS data integrity validation cycle*

6. ENSURING ONGOING DATA INTEGRITY

Unfortunately, continuously ensuring data integrity of a DRPS AIP does not end once the AIP has been written correctly to tape. Periodically, the tape(s) containing the AIP needs to be checked to uncover errors (i.e., bit flips) that may have occurred since the AIP was correctly written.

Fortunately, IBM LTO-5 and TS1140 tape drives can perform this check without having to stage the AIP to disk, which is clearly a resource intensive task—especially for an archive with a capacity measured in petabytes!

IBM LTO-5 and TS1140 drives can perform data integrity validation *in-drive*, which means a drive can read a tape and concurrently check the AIP logical block CRC and ECCs discussed above (C1, C2, and the original data CRC). Status is reported as soon as these internal checks are completed. And this is done without requiring any other resources!

Clearly, this advanced capability enhances the ability of DRPS to perform periodic data integrity validations of the entire archive more frequently, which will facilitate the correction of bit flips and other data errors.

7. LOOKING TO THE FUTURE

StorageGRID provides an HTTP API that automatically returns its SHA-1 hash values when called, but this API is not used at the present time because Rosetta currently only writes to permanent Network File System (NFS) storage using POSIX commands.

As a result of collaboration between Ex Libris and the Church History digital preservation team, the next version of Rosetta (3.1) will expose the Rosetta storage handler component as a Rosetta plugin. This will enable Rosetta to integrate with storage systems other than NFS, such as Amazon S3, storage systems which support CDMI (Cloud Data Management Interface), and others. The enhancement significantly expands Rosetta's reach into modern distributed file systems.

Ex Libris has committed to the Church a Rosetta plugin that will utilize the StorageGRID HTTP API and thus eliminate the need for the ICS-developed web service mentioned previously. This will provide a more elegant DRPS solution to fixity information hand off between Rosetta and StorageGRID.

As the size of the DRPS digital archive continues to grow, the need for increased Rosetta scalability is ever present. Fortunately for the Church, Ex Libris has been proactive in meeting its needs.

Subsequent to the original Rosetta scalability work mentioned earlier that was performed by the Church and Ex Libris together, significant improvements have been integrated to enhance Rosetta robustness and scalability.

For example, to fully leverage modern multicore processor technologies, a series of concurrent processing techniques have been implemented in Rosetta. Multi-threading is a programming and execution model that provides developers with a useful abstraction of concurrent execution. When applied to a single process, multi-threading permits parallel execution on a multiprocessor system, and also increases fault tolerance of the system.

Managing a concurrent flow has its challenges, however, since operations exclusivity and timing need to be continuously considered. To preclude errors, Java Messaging Service (JMS) was employed in Rosetta, allowing communications between different components of the distributed application to be loosely coupled, asynchronous, and reliable [3]. This enhancement provides robustness and fault tolerance, and guarantees that no work is lost.

Additional processing enhancements were implemented by using symbolic links for files during ingest and operational processes. These enhancements remove the need for copying files from one temporary location to another, thereby reducing I/O and improving network utilization as well as data integrity.

Large file ingest processing was also improved by incorporating DROID 6 file identification during the SIP validation stage. DROID 6 is substantially more efficient at identifying file formats of large files since it uses offsets to locate the file signature, and thus avoids a full scan of the entire file.

8. CONCLUSION

By working collaboratively with Ex Libris and utilizing advanced tape drives plus the sophisticated data integrity features of StorageGRID, the Church of Jesus Christ of Latter-day Saints has been able to advance the state of the art of data integrity and long term preservation in a rapidly growing digital preservation archive.

9. REFERENCES

- [1] CCSDS 650.0-B-1BLUE BOOK, “Reference Model for an Open Archival Information System (OAIS),” Consultative Committee for Space Data Systems (2002)
- [2] Private conversation with Sam Gustman (CTO) at the USC Shoah Foundation Institute August 19, 2009
- [3] <http://www.oracle.com/technetwork/java/jms/index.html>

Formats over Time: Exploring UK Web History

Andrew N. Jackson
The British Library
Boston Spa, Wetherby
West Yorkshire, LS23 7BQ, UK
Andrew.Jackson@bl.uk

ABSTRACT

Is software obsolescence a significant risk? To explore this issue, we analysed a corpus of over 2.5 billion resources corresponding to the UK Web domain, as crawled between 1996 and 2010. Using the DROID and Apache Tika identification tools, we examined each resource and captured the results as extended MIME types, embedding version, software and hardware identifiers alongside the format information. The combined results form a detailed temporal format profile of the corpus, which we have made available as open data. We present the results of our initial analysis of this dataset. We look at image, HTML and PDF resources in some detail, showing how the usage of different formats, versions and software implementations has changed over time. Furthermore, we show that software obsolescence is rare on the web and uncover evidence indicating that network effects act to stabilise formats against obsolescence.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Information filtering, Selection process*; H.m [Information Systems]: Miscellaneous

1. INTRODUCTION

In order to ensure that our digital resources remain accessible over time, we need to fully understand the software and hardware dependencies required for playback and reuse. The relationship between bitstreams and the software that makes them accessible is usually expressed in terms of data ‘format’ - instead of explicitly linking individual resources to individual pieces of software, we attach identifiers like file extensions, MIME types and PRONOM IDs to each and use that to maintain the link. These identifiers can also be attached to formal format specifications, if such documentation is available.

Successful digital preservation therefore requires us to fully understand the relationship between data, formats, software

and documentation, and how these things change over time. Critically, we must learn how formats become obsolete, so that we might understand the warning signs, choices and costs involved. This issue, and the arguments around the threat of obsolescence, can be traced back to 1997, when Rothenburg asserted that “Digital Information Lasts Forever—Or Five Years, Whichever Comes First.” [1]. Fifteen years later, Rothenberg maintains that this aphorism is still apt [2]. If true, this implies that all formats should be considered brittle and transient, and that frequent preservation actions will be required in order to keep our data usable. In contrast, Rosenthal maintains that this is simply not the case, writing in 2010 that “when challenged, proponents of [format migration strategies] have failed to identify even one format in wide use when Rothenberg [made that assertion] that has gone obsolete in the intervening decade and a half.” [3]. Rosenthal argues that the network effects of data sharing act to inhibit obsolescence and ensure forward migration options will arise. Similarly, Rothenburg remains skeptical of the common belief that different types of content are normalising on HTML5 and so reducing the number of formats we need to address [2]. If these assertions are true, then format migration or emulation strategies become largely unnecessary, leaving us to concentrate on storing the content and simply making use of the available rendering software.

The fact that the very existence of software obsolescence remains hotly disputed therefore undermines our ability to plan for the future. To find a way forward, we must examine the available evidence and try to test these competing hypotheses. In this paper, we begin this process by running identification tools over a suitable corpus, so that we can use the resulting format profile to explore what happens when formats are born, and when they fade away. Working in partnership with JISC and the Internet Archive (IA), we have been able to secure a copy of the IA web archives relating to the UK domain, and host it on our computer cluster. The collection is composed of over 2.5 billion resources, crawled between 1996 and 2010, and thus gives us a sufficiently long timeline over which some reasonable conclusions about web formats might be drawn.

Determining the format of each resource is not easy, as the MIME type supplied by the originating server is often malformed [4]. Instead, we apply two format identification tools to the content of each resource - DROID and Apache Tika. Both use internal file signature (or ‘magic numbers’) to identify the likely format of each bitstream, but

differ in coverage, complexity and granularity. In particular, DROID tuned to determine different versions of formats, while Apache Tika returns only the general format type, but augments it with more detailed information gleaned from parsing the bitstream. Thus, by combining both sets of results, we can come to a more complete understanding of the corpus. Furthermore, by comparing the results from the different identification tools, we can also uncover inconsistencies, problematic formats and weak signatures, and so help drive the refinement of both tools.

2. METHOD

The test corpus is called the JISC UK Web Domain Dataset (1996-2010), and contains over 2.5 billion resources harvested between 1996 and 2010 (with a few hundred resources dated from 1994), either hosted on UK domains, or directly referenced from resources hosted under ‘.uk’. This adds up to 35TB of compressed content held in 470,466 arc.gz and warc.gz files, now held on the a 50-node HDFS filesystem. As the content is hosted on this distributed filesystem, we are able to run a range of tools over the whole dataset in a reasonable time using Hadoop’s Map-Reduce framework.

Due to its prominence among the preservation community and the fine-grained identification of individual versions of formats, DROID was chosen as one of the tools. To complement this, we also chose to use the popular Apache Tika identification tool, which has been shown to have much broader format coverage [5]. Unfortunately, both tools required some modification in order to be used in this context. DROID was particularly problematic, and we were unable to completely extract the container-based identification system in a form that made it re-usable as a Map-Reduce task. However, the binary file format identification engine could be reused, and the vast majority of the formats that DROID can identify are based on using that code (and the DROID signature file it depends upon - we used signature file version 59). Herein, we refer to this as the ‘DROID-B’ tool. Both tools were run directly on the bitstreams, rather than being passed the URLs or responses in question, and so the identification was based upon the resource content rather than the name or any other metadata. For this first experimental scan, we decided to limit the identification process to the top-level resource associated with each URL and crawl time - archive or container formats were not unpacked.

In order to compare the results from DROID-B and Apache Tika with the MIME type supplied by the server, the identification results are normalised in the form of extended MIME Types. That is, where we know the version of a format as well as the overall MIME Type, we add that information to the identifier using a standard type parameter, e.g. “image/png; version=1.0”, corresponding to PUID fmt/11. In this way, extended MIME types can act as a bridge between the world of PRONOM identifiers and the standard identification system used on the web. Broad agreement between tools can be captured by stripping off the parameters, but their presence lets more detailed information be collected and compared in simple standard form. A number of formats also embed information about the particular software or hardware that was using in their creation - PDF files have a ‘creator’ and a ‘producer’ field, and many image formats have similar EXIF tags. As we are also in-

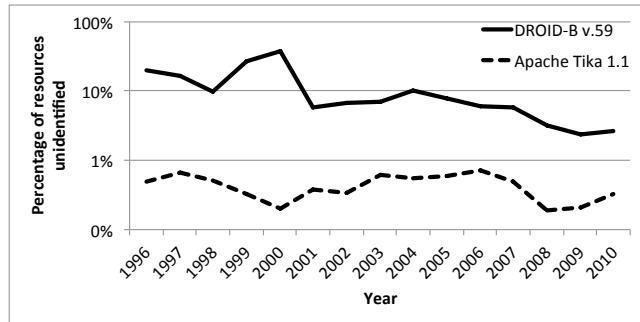


Figure 3.1: Identification failure rates for Apache Tika and DROID-B.

terested in the relationship between software and formats, we have attempted to extract this data and embed it in the extended MIME type as software and hardware parameters. The full identification process also extracted the year each resource was crawled, and combines this with the three different MIME types to form a single ‘key’. These keys were then collected and the total number of resources calculated for each. Overall, the analysis was remarkably quick, requiring just over 24 hours of continuous cluster time.

3. RESULTS

3.1 The Format Profile Dataset

The primary output of this work is the format profile dataset itself¹. Each line of this dataset captures a particular combination of MIME types (server, Apache Tika and DROID-B), for a particular year, and indicates how many resources matched that combination. For example, this line:

```
image/png image/png image/png; version=1.0 2004 102
```

means that in this dataset there were 102 resources, crawled in 2004, that the server, Tika and DROID-B all agreed have the format ‘image/png’, with the latter also determining the format version to be ‘1.0’. Due disagreements over MIME types and the number different hardware and software identifiers the overall profile is rather large, containing over 530,000 distinct combinations of types and year. Below, we document some initial findings drawn from this rich dataset, and we have made it available under an open licence (CC0) in the hope that others will explore and re-use it.

3.2 Comparing Identification Methods

3.2.1 Coverage & Depth

The identification failure rates for both tools are shown in Figure 3.1, as a percentage of the total number of resources from each year. Overall, Apache Tika has significantly lower failure rate than DROID-B - 1% versus around 10%. There also seems to be a significant downward trend in the DROID-B curve, which would indicate that DROID copes less well with older formats. However, initial exploration indicate that this is almost entirely due to the prevalence of pre-2.0 HTML, which was often poorly formed.

¹To download the dataset, see <http://dx.doi.org/10.5259/ukwa.ds.2/fmt/1>.

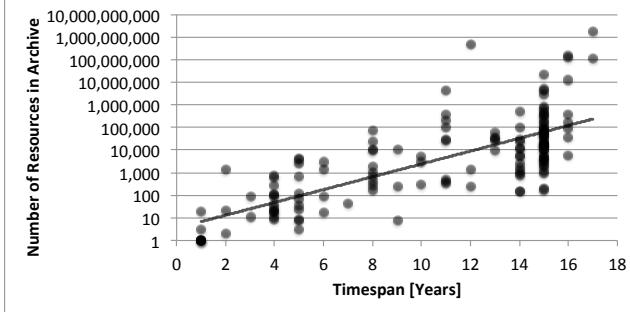


Figure 3.2: Number of resources of each format versus its lifespan. Formats identified using Apache Tika.

3.2.2 Inconsistencies

By comparing the simple MIME types (no parameters) we were able to compare the results from both tools, revealing 174 conflicting MIME type combinations. For example, some 2,957,878 resources that Apache Tika identified as ‘image/jpeg’ we identified as ‘image/x-pict’ by DROID. The PRONOM signature for this format is rather weak (consisting of a single byte value at a given offset) and can therefore produce a large number of false positives when run at scale². Another notable class of weak signatures correspond to text-based formats like CSS, JavaScript, and older or malformed HTML. Apache Tika appears to perform slightly better here - for example, the HTML signature is much more forgiving than the DROID-B signature.

More subtle inconsistencies arose for the Microsoft Office binary formats and for PDF. In the former case, a full implementation of DROID would probably be able to resolve many of the discrepancies. The picture for PDF is more complex. The results were mostly consistent, but DROID-B failed to recognise 1,340,462 resources that Apache Tika identified as PDF. This appears to be because the corresponding PRONOM signature requires the correct end-of-file marker (‘%EOF’) to be present, whereas many functional documents can be mildly malformed, e.g. ending with ‘%EO’ instead. Also, the results for PDF/A-1a and PDF/A-1b were not entirely consistent, with Tika failing to identify many documents that DROID-B matched, but matching a small number of PDF/A-1b documents that DROID missed. A detailed examination of the signatures and software will be required to resolve these issues.

3.3 Format Trends

As mentioned in the introduction, one of the core questions we need to understand is whether formats last a few years and then die off, or whether (on the web at least) network effects take over and help ensure formats survive. We start to examine this question by first determining the lifespan of each format - i.e. the number of years that have elapsed between a format’s first and last appearance in the archive. This lifespan is plotted against the number of resources that were found to have that format, such that young and rare

²Indeed, it appears that this signature has been removed from the latest version of the DROID binary signature file (version 60, published during preparation).

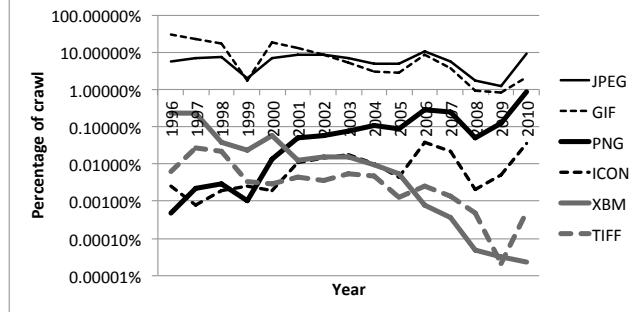


Figure 3.3: Selected popular image formats over time. Formats identified using Apache Tika.

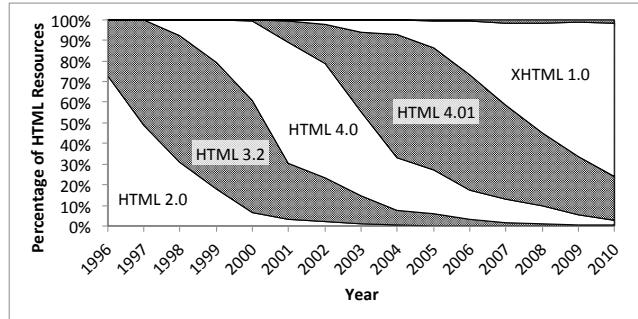


Figure 3.4: HTML versions over time. Formats identified using DROID-B.

formats appear in the bottom-left corner, whereas older and popular formats appear in the top-right, as shown in figure 3.2. Due to the extreme variation in usage between formats, the results are plotted on a logarithmic scale.

If popularity has no effect on lifespan, we would expect to see a simple linear trend - i.e. a format that has existed for twice as long as another would be found in twice as many documents. Due to the logarithmic vertical axis of figure 3.2, would be shown as a sharp initial increase followed by an apparent plateau. However, in the presence of network effects we would expect a much stronger relationship, and indeed this is what we find - a format that has been around longer is exponentially more common than younger formats (an exponential fit appears as a straight line in figure 3.2). A large number of formats have persistent for a long time (47 formats have been around for 15 years), and that since 1997, roughly six new formats have appeared each year while fewer have been lost (roughly 2 per year). While this confirms the presence of the network effects Rosenthal proposed, proving that these formats are more resilient against obsolescence will require a deeper understanding of obsolescence itself.

As a first step in that direction, we examine how format usage changes over time. Figure 3.3 shows the variation in usage of some of the most common image formats. Unsurprisingly, JPEG has remained consistently popular. In contrast, the PNG and ICO formats have become more popular over time, and the GIF, TIFF and XBM formats have decreased in popularity, with the drop in usage of the XBM format being particularly striking.

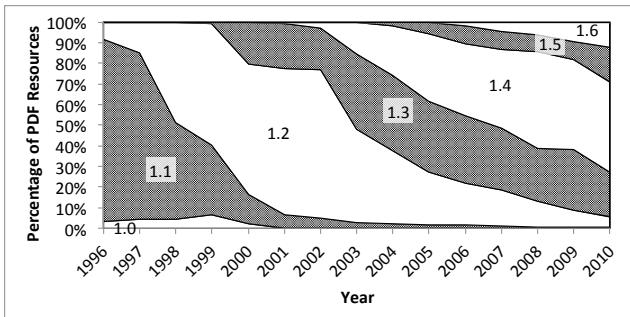


Figure 3.5: PDF versions over time. Formats identified using DROID-B and Apache Tika.

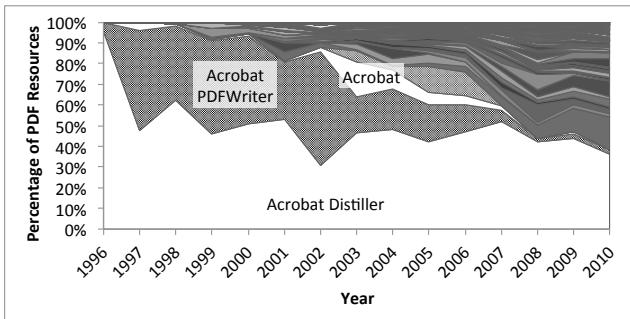


Figure 3.6: PDF software identifiers over time. Formats and software identified using Apache Tika.

3.4 Versions & Software

Figures 3.4 and 3.5 show how the popularity of various versions of HTML and PDF has changed over time. In general, each new version grows and dominates the picture for a few years, before very slowly sinking into obscurity. Thus, while there were just two active versions of HTML in 1996 (2.0 and 3.2), all six were still active in 2010. Similarly, there were three active versions of PDF in 1996 (1.0-1.2) and eleven different versions in 2010 (1.0-1.7, 1.7 Extension Level 3, A-1a and A-1b, with 1.2-1.6 dominant). In general, it appears that format versions, like formats, are quick to arise but slow to fade away.

Finally, figure 3.6 shows the popularity of different software implementations over time and the dominance of the Adobe implementations (although later years have seen an explosion in the number of distinct creator applications, with over 2100 different implementations of around 600 distinct software packages). Similarly, the JPEG data revealed over 1900 distinct software identifiers and over 2100 distinct hardware identifiers. We speculate that the number of distinct implementations can be taken as an indicator for the maturity, stability and degree of standardisation of a particular format, although more thorough analysis across more formats would be required to confirm this.

4. CONCLUSIONS

We have made a rich dataset available, profiling the format, version, software and hardware data from large web archive spanning almost one and a half decades. Our initial analysis supports Rosenthal's position; that most formats last much

longer than five years, that network effects to appear to stabilise formats, and that new formats appear at a modest, manageable rate. However, we have also found a number of formats and versions that are fading from use, and these should be studied closely in order to understand the process of obsolescence. Furthermore, we must note that every corpus contains its own biases, such as crawl size limits or scope parameters³. Therefore, we recommend that similar analyses be performed on a wider range of different corpora in order to attempt to confirm these trends.

We used two different tools (DROID-B and Apache Tika) that perform the essentially the same task (format identification), and ran them across the same large and varied corpus. In effect, each can be considered a different ‘opinion’ on the format, and by uncovering the inconsistencies and resolving them, we can improve the signatures and tools in a very concrete and measurable way, and more rapidly approach something like a ‘ground truth’ corpus for format identification.

Future work will examine whether the underlying biases of the corpus can be addressed, whether we can reliably identify resources within container formats, and whether the raw resource-level data can be made available. This last point would allow many more format properties to be exposed and make it easier to resolve inconsistent results by linking back to the actual resources.

5. ACKNOWLEDGMENTS

This work was partially supported by JISC (under grant DI-INN06) and by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

6. REFERENCES

- [1] Rothenberg, Jeff; (1997) “Digital Information Lasts Forever—Or Five Years, Whichever Comes First.” RAND Video V-079
- [2] Rothenberg, Jeff; (2012) “Digital Preservation in Perspective: How far have we come, and what’s next?” Future Perfect 2012, New Zealand (Archived by WebCite® at <http://www.webcitation.org/68OuQxEHj>)
- [3] Rosenthal, David S.H.; (2010) “Format obsolescence: assessing the threat and the defenses”, Library Hi Tech, Vol. 28 Iss: 2, pp.195 - 210
- [4] Clausen, Lars R.; (2004) “Handling file formats” (Archived by WebCite® at <http://www.webcitation.org/68PyiaA9w>)
- [5] Radtisch, Markus; May, Peter; Askov Blekinge, Asger; Møldrup-Dalum, Per; (2012) “SCAPE Deliverable D9.1: Characterisation technology, Release 1 & release report.” (Archived by WebCite® at <http://www.webcitation.org/68OttmVnn>)

³Even the crawl time itself can be quite misleading, as a newly discovered resource may have been created or published some years before

From cataloguing to digital curation: the role of libraries in data exchange

Susan K. Reilly

LIBER

Koninklijke Bibliotheek

Den Haag

+31 (0)703140160

susan.reilly@kb.nl

ABSTRACT

This paper describes the work of the Opportunities for Data Exchange (ODE) project, a project funded by the European Commission under Framework Programme 7. This project investigates issues surrounding data preservation, reuse and exchange from both sociological and technical view points.

Led by the European Organisation for Nuclear Research (CERN), the project has sought out stories of success and honorable failures. It has also brought together representatives of key stakeholder in the data preservation and sharing landscape. This has enabled dialogue between these stakeholder in order to identify opportunities for researchers, publishers and libraries to play their part in data exchange.

The growing need for research data preservation and curation services, the linking of data to publications, and increasing awareness of the potential of data sharing for innovation, presents a major opportunity for libraries to redefine their roles and embed themselves in the research process. In November 2011 ODE surveyed the 420 plus LIBER member libraries to establish what demand from researchers libraries are experiencing for support in data exchange, what roles they need to fulfill, and what new skills they need to develop and how. The results clearly emphasised the importance of the development of the role of the library in digital curation.

Categories and Subject Descriptors

H.3.5 [Online Information Services]: Data sharing

General Terms

Human Factors

Keywords

Research data sharing, digital curation, libraries

1. INTRODUCTION

Funded by the European Commission under Framework Programme 7, the Opportunities for Data Sharing¹ (ODE) project's aim was to identify, collate, interpret and deliver evidence of emerging best practices in sharing, re-using, preserving and citing data, the drivers for these changes and barriers impeding progress.

This was done in forms suited to its target audiences/stakeholders of policy makers, funders, infrastructure operators, data centres, data providers and users, libraries and publishers.

The aim of the project has been to:

- Enable operators, funders, designers and users of national and pan-European e-Infrastructures to compare their vision and explore shared opportunities
- Provide projections of potential data re-use within research and educational communities in and beyond the ERA, their needs and differences
- Demonstrate and improve understanding of best practices in the design of e-Infrastructures leading to more coherent national policies
- Document success stories in data sharing, visionary policies to enable data re-use, and the needs and opportunities for interoperability of data layers to fully enable e-Science
- Make that information available in readiness for FP8

Within this context, the stakeholder representatives in the project have worked together to engage and raise the profile of data sharing, re-use and preservation as an issue with each of our communities and to undertake further, in-depth, investigation into the issues raised.

LIBER, the Association for European Research Libraries, represents over 420 research libraries from across Europe. Through LIBER, ODE has engaged research libraries in Europe in the dialogue surrounding data exchange on issues such as linking data to publications, best practice in

¹ www.ode-project.eu

data citation and, subsequently, exploration of the role of libraries in supporting data exchange.

2. IDENTIFYING OPPORTUNITIES

The data deluge and its implications has been explored by the High level expert Group on Scientific Data in the Riding the Wave report². The report outlines the need for the development of an international framework for a collaborative data infrastructure. This framework is described as broad conceptual framework which outlines how stakeholders interact with the system, including a multitude of actors, with provisions for data curation at every layer.

ODE goes some way in exploring this interaction through the identification of common issues, drivers and barriers in data exchange. One of the ways in which these are explored is through an analysis of the impact that data sharing, re-use and preservation is having on scholarly communication³. The aim of this analysis is to identify incentives for researchers and other stakeholders that will help to optimise the take-up of future e-Infrastructures.

One of the key areas of opportunity in terms of exploiting and proving the value of data exchange is scholarly communications. The opportunity to share and interact with research data is changing the face of scholarly communication and creating new opportunities and challenges for researchers, publishers and libraries. Publishing the underlying data of an article creates greater transparency and potentially further research, but it must also be in the interest of the data creator to publish and the data much be published in a manner which is sustainable. Three areas have been examined by ODE in relation to scholarly communications: linking data to publications, best practice in data citation, and the evolving role of libraries.

The findings of the exploration of linking data to publications were published in a report, which sought to reveal opportunities for supporting a more connected and integrated scholarly record. Four perspectives were considered, those of the researcher, who generates or reuses primary data, publishers, who provide the mechanisms to communicate research activities, and libraries & data centers, who maintain and preserve the evidence that underpins scholarly communication and the published record.

Before identifying opportunities it is necessary to look at the different layers (fig.1.) of data publication and identify issues associated with each layer.

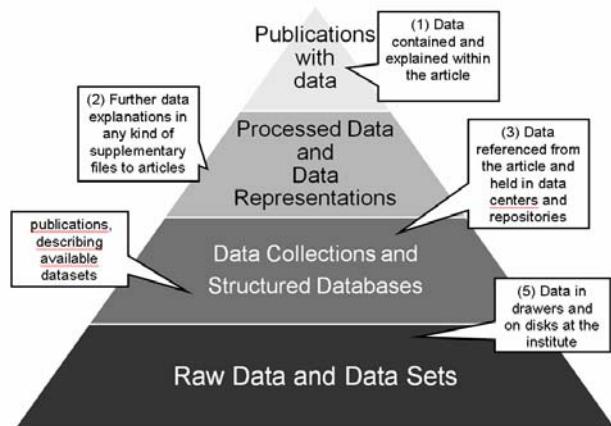


Figure 1. ODE Data Publication Pyramid

Each layer presents different challenges and opportunities e.g. one of the challenges presented by bottom layer of raw data is to encourage researchers to deposit their data in a sustainable infrastructure. The report identified opportunities for all three groups in seven key areas:

1. Libraries have the opportunity to support **availability** by helping researchers make their data available and also providing search services for data.
2. Through the provisions of support for best practice in managing data they can support **findability**.
3. As experts in metadata they can support **interpretability** through the provision of, and training in, metadescritions.
4. By advising on the availability of subject archives and licensing for reuse libraries can help work towards ensuring the **reusability** of research data.
5. By encouraging best practice in citations through the provision of guidance and training, and through the use of persistent identifiers for data sets libraries play a role in improving the **citability** of data sets.

² The High Level Expert group on Scientific Data (2010), Riding the Wave, <http://www.cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>

³ Reilly et al. (2011) ODE report on the integration of data and publications:<http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/10/ODE-ReportOnIntegrationOfDataAndPublications.pdf>

6. Libraries can also take on some responsibility for the **curation** of data and provision of training on data curation

7. Contribute to the long term **preservation** of data by advocating for good data management practices and the archiving of data.

In essence, developments in linking data to publications and, more broadly, data exchange presents libraries with the opportunity to redefine their roles and become more embedded in the research process. Libraries should not underestimate their role as advocates for data sharing and for best practice in data management. This examination of linking data to publications also points to the fact that libraries are well placed to provide support for data curation across the layers of a collaborative data infrastructure.

3. REDEFINING ROLES

These seven areas of opportunity were presented to a group of librarians during a workshop at the 2011 LIBER Annual Conference in Barcelona. What emerged from this workshop was a very clear need for libraries to clarify their roles in relation to data exchange and the opportunities identified. Furthermore there is a need to understand these roles so that this can inform the identification of existing skills to be built on and new skills to be developed. The libraries were in consensus that they were in a strong position to address fragmentation in curation and archiving but there were doubts surrounding whether they were equipped to take decisions regarding what research data should be curated and archived or even what their role should be in making these decisions.

4. SURVEY OF RESEARCH LIBRARIES

The workshop established that libraries are keen to engage in data exchange but that further exploration of the types of roles libraries should play in this was needed. To follow up on this a survey was sent out to all 430 libraries in the LIBER network. The spectrum of libraries within the LIBER network covers national and state libraries, as well as university libraries and research institutes.

The survey was designed to gather evidence on the current and expected roles of libraries in regard to data management in order to prescribe steps for the evolution of these roles. This has been done through gathering answers from libraries related to the following questions:

1. What is the perceived demand from researchers for support for data management from libraries?

2. In what areas does this demand exist?

3. What support is currently in place?

4. What skills are needed to meet the demand for support?

In total 110 responses were gathered, from a mailing to LIBER members that reaches approximately 800 people (response rate 13 %). Additional responses were gathered from a dozen internationally recognized leading libraries (experts) in the field of data management support from the US and Australia. As these select few were already active in the field their responses were meant to form a benchmark.

4.1 Survey Results

The responses to the survey make it clear that librarians regard their involvement in support for research data exchange as a new and important role. For the majority, the service level is still rather low, but librarians also appear keen to develop themselves in the area of data management, archiving and curation as well as in helping their researchers find data. The survey received a response rate of nearly 20% and so can be judged as representative of the state of play across research libraries in Europe.

4.2 Demand for Support

81% of the respondents reported a demand for data management support. Considering that the response came for the broad spectrum of European research libraries and not just large university libraries this is quite a high figure. What came out most clearly in the survey results is that libraries are nowhere near meeting the perceived demand for support (fig.2). The area where most demand was perceived was for archiving data. 80% of respondents perceived a demand in this area, yet only 41% of these respondents actually provide any sort of support services for the archiving of data.

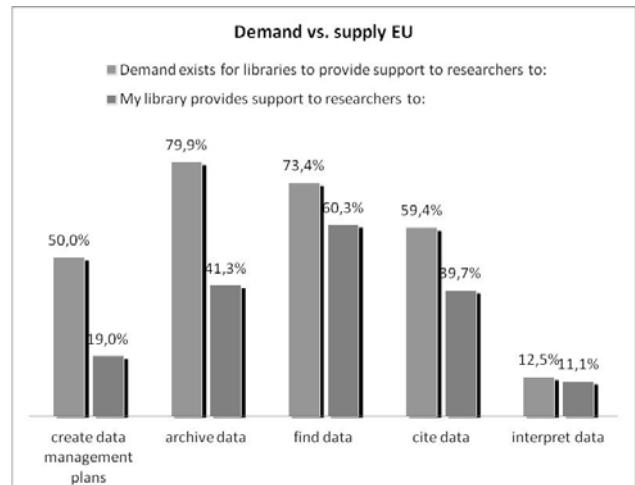


Figure 2. Demand v. Supply

When asked who should be responsible for the selection of data for archiving, respondents saw this responsibility as lying with researchers, followed by data librarians, librarians and others. Interestingly, the Expert libraries differ with this opinion in that they all agreed that only the researchers should be responsible for the selection, and no one else. There were qualifiers to this response stating that libraries should work with researchers to do this.

The majority of respondents (66%) also reported that their institutions did not have strategies in place for the preservation of research data, which is a worrying gap.

Only 6% of libraries had an internal archiving system for archiving research data and furthermore only 10% cooperated with a disciplinary data archive. That only 6% of respondent libraries have an archive for research data is not necessarily worrying. In many cases it is preferable and, arguably, more sustainable, to encourage researchers to use a disciplinary archive. What is of concern, is the fact that so few libraries seem to be collaborating with disciplinary archives. This is not just worrying from a library perspective, but also from the perspective of those funding such infrastructures. To ensure successful uptake and exploitation, such research infrastructures should be working with stakeholders such as libraries to help them to encourage and provide support to their researchers to use these collaborative infrastructures.

An encouraging figure is that a greater number of libraries are employing their traditional cataloging related skills when it comes to making sure that data remains interpretable and reusable. 39% of respondents report that they use metadata to ensure this. This shows that libraries are already adapting their existing skills to meet the increasing demand for data management support.

4.3 Developing Skills

The two key areas where skills need to be developed are IT and data curation. Responses from Europe showed that IT skills were seen to be the most important area for skills development. On the other hand, the Expert libraries strongly prioritised digital curation as an area for the development of skills. For them, IT skills came in 4th place in terms of priority. It may be that experience shows that IT skills are not as important as perceived for libraries who are actively involved in data management support.

The best means of developing all such skills, according to the libraries, is through the provision of continuing professional development. During times when budgets are contracting it is not realistic to expect to be able to recruit new skill sets externally. Instead libraries must, where possible, invest in developing the skills of existing staff. This solution may not be entirely sufficient, particularly when it comes to the need for subject specific expertise. Subject specific expertise was prioritized by 88% of Expert

libraries and 67% of European libraries. Ultimately, the demand for such expertise may lead to new approaches to the professional education and recruitment of librarians.

5. CONCLUSION

It is clear that there is an opportunity and demand for libraries to provide support in digital curation. What has not been so clear is what exactly the nature of libraries' role should be in this. Perhaps wisely, libraries have realized that they can not take on full responsibility for the curation of research data. Researchers must be involved in the selection of data for archiving. If researchers are to be solely responsible for this, then libraries should begin to consider how they can support researchers to make these decisions?

On the other hand, the current support for digital curation is not sufficient. Libraries can apply their traditional skills to this area but they must also invest in developing new skills to meet demand for support and to avoid what could be a very regrettable missed opportunity. A start would be to put strategies for the preservation of research data in place, these strategies might involve the establishment of an internal archive that supports the persistent identification of data sets or they could be as simple as collaborating with disciplinary archives on behalf of their own research communities.

6. REFERENCES

- [1] ParseInsight Survey (2009) http://www.parse-insight.eu/downloads/PARSE-Insight_D4-GapAnalysisFinalReport.pdf
- [2] Reilly et al. (2011) ODE report on the integration of data and publications: <http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/10/ODE-ReportOnIntegrationOfDataAndPublications.pdf>
- [3] Survey Report PARSE.Insight:http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf
- [4] Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, et al. (2011) Data Sharing by Scientists: Practices and Perceptions. PLoS ONE 6(6): e21101. doi:10.1371/journal.pone.0021101
- [5] The High Level Expert group on Scientific Data (2010), Riding the Wave, <http://www.cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>

The Community-driven Evolution of the Archivematica Project

Peter Van Garderen
President, Artefactual Systems, Inc.
202-26 Lorne Mews
New Westminster, BC, Canada
1.604.527.2056
peter@artefactual.com

Courtney C. Mumma
Systems Archivist, Artefactual Systems Inc.
202-26 Lorne Mews
New Westminster, BC, Canada
1.604.527.2056
courtney@artefactual.com

ABSTRACT

In this paper, we discuss innovations by the Archivematica project as a response to the experiences of early implementers and informed by the greater archival, library, digital humanities and digital forensics communities. The Archivematica system is an implementation of the ISO-OAIS functional model and is designed to maintain standards-based, long-term access to collections of digital objects. Early deployments have revealed some limitations of the ISO-OAIS model in the areas of appraisal, arrangement, description, and preservation planning. The Archivematica project has added requirements intended to fill those gaps to its development roadmap for its micro-services architecture and web-based dashboard. Research and development is focused on managing indexed backlogs of transferred digital acquisitions, creating a SIP from a transfer or set of transfers, developing strategies for preserving email, and receiving updates about new normalization paths via a format policy registry (FPR).

General Terms

Documentation, Performance, Design, Reliability, Experimentation, Security, Standardization, Theory, Legal Aspects.

Keywords

archivematica, digital preservation, archives, OAIS, migration, formats, PREMIS, METS, digital forensics, agile development, open-source, appraisal, arrangement, description, acquisition

1. INTRODUCTION

The ISO 14721-OAIS Reference Model [1] gave the archives community a common language for digital archives architectures. One such architecture is the Archivematica suite of tools which

was based on an extensive requirements analysis of the OAIS functional model [2]. The Archivematica project is nearing its first beta release. Project partners and independent implementers have been testing alpha releases using real-world records. These activities have identified some OAIS requirement gaps for digital archives systems.

The project has found that, while it serves as an excellent foundation and framework for long-term preservation strategies, the OAIS model proves inadequate to address some functions unique to archives. In particular for the areas of appraisal, arrangement, description, and preservation planning there were clear gaps between the model and the way that archivists actually process records. The Archivematica project has added requirements to its development roadmap to fill those gaps in its micro-services architecture and web-based dashboard. Other research and development is focused on managing a backlog of indexed digital acquisitions, creating a Submission Information Package (SIP) from a transfer or set of transfers, developing strategies for preserving email, and receiving updates about new normalization paths via a format policy registry (FPR).

2. ABOUT THE ARCHIVEMATICA PROJECT

The Archivematica system uses a micro-services design pattern to provide an integrated suite of free and open-source software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model [3]. It allows archivists and librarians to process digital transfers (accessioned digital objects), arrange them into Submission Information Packages (SIPs), apply media-type preservation plans and create high-quality, repository-independent Archival Information Packages (AIPs). Archivematica is designed to upload Dissemination Information Packages (DIPs) containing descriptive metadata and web-ready access copies to external access systems such as DSpace, CONTENTdm and ICA-AtOM. Users monitor and control the micro-services via a web-based dashboard.

A thorough use case and process analysis identified workflow requirements to comply with the OAIS functional model. Through deployment experiences and user feedback, the project has expanded beyond OAIS requirements to address analysis and arrangement of transfers into SIPs and allow for archival appraisal at multiple decision points. The Archivematica micro-

services implement these requirements as granular system tasks which are provided by a combination of Python scripts and one or more of the free, open-source software tools bundled in the Archivematica system.

Archivematica uses METS, PREMIS, Dublin Core and other recognized metadata standards. The primary preservation strategy is to normalize files to preservation and access formats upon ingest when necessary (for example, when the file is in a format that is proprietary and/or is at risk of obsolescence). The media type preservation and access plans it applies during normalization are based on format policies derived from an analysis of the significant characteristics of file formats [4]. The choice of access formats is based on the ubiquity of viewers for the file format as well as the quality of conversion and compression. Archivematica's preservation formats are all open standards [5]. Additionally, the choice of preservation and access formats is based on community best practices and availability of open-source normalization tools.

Archivematica maintains the original files to support future migration and emulation strategies. However, its primary preservation strategy is to normalize files to preservation and access formats upon ingest. The default normalization format policies can be edited and disabled.

All of the software, documentation and development infrastructure are available free of charge and released under AGPL3 and Creative Commons licenses to give users the freedom to study, adapt and re-distribute these resources as best suits them. Archivematica development is led by Artefactual Systems, a Vancouver based technical service provider that works with archives and libraries to implement its open-source solutions as part of comprehensive digital preservation strategies. All funding for Archivematica development comes from clients that contract Artefactual's team of professional archivists and software developers to assist with installation, integration, training and feature enhancements. The majority of Archivematica users take advantage of its free and open-source license without additional contracting services.

3. ACQUISITION AND BACKLOG MANAGEMENT

Early implementers of the Archivematica suite of tools have consistently struggled with the mechanics of acquiring digital materials. Analogue records are delivered to the repository or are picked up from the donor's storage location, but digital acquisition can be more varied. Digital materials can arrive via digital transfer over a network such as email, FTP or shared directories. The archives may have to send an archivist to acquire the digital materials onsite, and even then, there are several options for acquisition including pickup, copying, or imaging. Depending on the type of acquisition, should the archivist photograph the condition of the materials in their original location? What steps must be taken to ensure that the digital objects copied or imaged retain their integrity during transfer to the archives? Finally, when digital materials are donated to the archives onsite, how do processes differ from pickup and digital network transfer?

Archivists who deal primarily with analogue materials are well accustomed to the need to maintain a backlog. Acquisitions regularly occur for which there are limited or no resources to process them immediately. For this reason, it is imperative that

the archives achieve a minimum level of control over the material so that it can be tracked, managed, prioritized and, if necessary, subjected to emergency preservation actions.

Archivematica runs through a set of transfer actions in the dashboard to establish initial control of the transfer. It verifies that the transfer is properly structured or structures it if necessary. Then, it assigns a unique universal identifier (UUID) for the transfer as a whole and both a UUID and a sha-256 checksum to each file in its /objects directory. Next, Archivematica generates a METS.xml document that captures the original order of the transfer and that will be included in any SIP(s) generated from this transfer. Any packaged files are unzipped or otherwise extracted, filenames are sanitized to remove any prohibited characters, and file formats are identified and validated. Finally, technical metadata is extracted from the files and the entire transfer content and metadata is indexed. At this point in the process, the transfer is ready to be sent to a backlog storage location that should be maintained in much the same way as the archival storage. The transfer is ready for future processing. These features will be added and evaluated in forthcoming releases of the Archivematica software.

4. ARRANGEMENT AND DESCRIPTION

Once an archive is ready to process one or more digital acquisitions, the next challenge comes from making a SIP from disparate parts of an acquisition. For example, in a situation in which an acquisition arrives on multiple digital media, the archive may have accessioned transfers from each media type and/or broken a very large hard drive into two or more transfers. Presumably, archivists will want their SIPs to be formed so that the resultant AIPs and DIPs conform to some level of their archival description, so SIP content could derive from one or more transfers or parts of transfers.

Arrangement and description do not neatly occur at one specific point during processing. Archivists arrange and describe analogue records intermittently. Arrangement is based upon the structure of the creator's recordkeeping system, inherent relationships that reveal themselves during processing and compensations made to simplify managing records and/or providing access. Archivists document their arrangement decisions and add this information, along with additional descriptive information gathered about the records during processing, to the archival description. Further, documentation of arrangement decisions and actions supports respect des fonds by preserving information about original order. Digital records must be arranged and described in order to effectively manage and provide access to them. Analogue functionality is very difficult to mimic in a digital preservation system such as Archivematica, because any interaction that allows for analysis of the records can result in changing original order and metadata associated with the records.

The OAIS model assumes that a digital archive system receives a fully formed SIP. However, this is often not the case in practice. Early Archivematica implementers were often manually compiling SIPs from transfers in the Thunar file browser bundled with the system. After transfer micro-services are completed successfully, Archivematica allows transfers to be arranged into one or more SIPs or for one SIP to be created from multiple transfers. The user can also re-organize and delete objects within the SIP(s). The original order of the transfer is maintained as its own structMap section in the transfer METS

file, a copy of which is automatically added to each SIP. Additionally, the archivist can use dashboard functionality to add basic descriptive metadata to the SIP at this point, including information about rights and restrictions.

The Archivematica project is now working on the ability to call up a transfer into a file browser interface in the dashboard's Ingest tab, examining its contents and forming it into SIPs for processing (See Figure 1).

review massive sets of digital records and compile selections from them as evidence. Clearly, the set of records presented as evidence must be verifiably authentic. Since archives are held to the same standards of authenticity there is much to be learned from the digital forensics field, which for over thirty years has been developing tools for processing evidence that guarantees its acceptance in courts. Such tools allow for auditing an investigator's actions, recording information about the set of

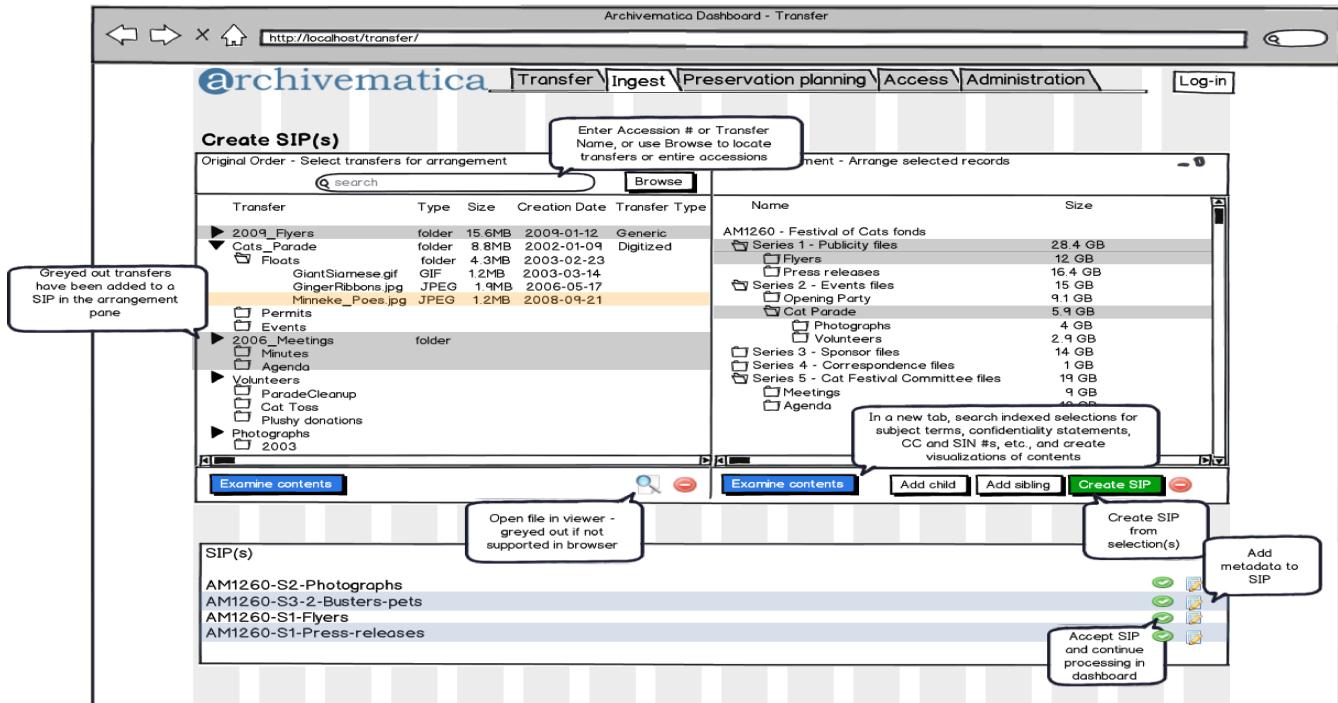


Figure 1. Create SIP dashboard interface mockup.

Much of the inspiration for such an interface came from digital forensics software and the Curator's Workbench at the University of North Carolina, Chapel Hill. The UNC Libraries had developed Curator's Workbench [6], a tool that, among other things, allows for arrangement of digital records without losing the original order. The Archivematica team considered including the tool in their suite, but because of concerns about integration and ongoing support, they opted instead to mimic its arrangement functionality. Archivematica's 0.8 alpha release uses the Xubuntu file browser Thunar to arrange records and METS to keep a record of the original order within each SIP formed from a transfer. In future releases, the METS file is still generated while file browser functionality has been moved to the dashboard. Future developments could see expanded METS and/or PREMIS profiles that includes information about selection actions undertaken during SIP creation and at the various appraisal stages.

The limitations for analyzing and forming SIPs using only the file browser were clear in earlier Archivematica releases. Transfers may contain restricted material, passwords, personal information or other content that is unsuitable for continued preservation. For insight into this problem, the project explored the possibilities of using digital forensics techniques. Digital forensics experts must

records and its origin while adding descriptive metadata and grouping portions of the set into discrete evidence packages, indexing and examining the file system structure and contents, and ensuring integrity. Many of the software tools used by digital forensics experts are proprietary, but in recent years open source tools have been developed to perform the same functions.

Despite their availability, open source digital forensics tools can be difficult to understand by non-experts. Serendipity's role in open source software development cannot be overstated. Just when Archivematica's systems analysts realized that they could not possibly decipher the entire canon of digital forensics software in time for the next release, digital humanities scholars and archivists in the United States were conceptualizing the BitCurator Project. From the BitCurator website [7]: "The BitCurator Project is an effort to build, test, and analyze systems and software for incorporating digital forensics methods into the workflows of a variety of collecting institutions." Artefactual Systems is closely involved with the BitCurator Project, with its president, Peter Van Garderen, on the Development Advisory Group and Courtney Mumma, systems analyst, on the Professional Experts Committee. Ideally, BitCurator will result in a set of open source tools that allow for arrangement,

description and other valuable functionalities that integrate well into the Archivematica suite.

Since open source digital forensics tools for archivists like BitCurator are not yet ready to be integrated into the Archivematica suite, the team looked for other ways to provide the necessary services to satisfy their workflows. One possible solution is using Apache Tika [8] and ElasticSearch [9] to index and search transfers in a dashboard file browser window to determine which part(s) to include in the SIP and to create visualizations of the transfer and SIP contents.

Requirements for future releases include indexing and reporting on all text content, file embedded metadata and file formats. Using Tika, ElasticSearch and other tools, Archivematica will provide keyword and pattern matching for privacy/security sensitive information (e.g. social insurance numbers/social security numbers, credit card numbers, email addresses and security keywords) and reports of such things as PDFs that have not been OCR'ed, password protected and encrypted files and duplicates with their full file paths. Reports for all of these indexing requirements will be available via the Examine Contents windows, accessible from the Create SIP browser window in the Ingest tab of the dashboard.

The Examine Contents reports will include a search box for the indexed transfer content, general information about the transfer or selected file group (e.g. number of files, size, name, UUID, and accession number), a pie graph visualization showing file type distribution overall and a bargraph visualization showing file type by folder and ordered by size. Clickable links will open to sub-reports on all contents of a specified format in context of the entire transfer, duplicates with their locations, privacy and confidentiality keywords and numbers and password protected files with their distribution across the entire contents visualized as a graph (See Figure 2).

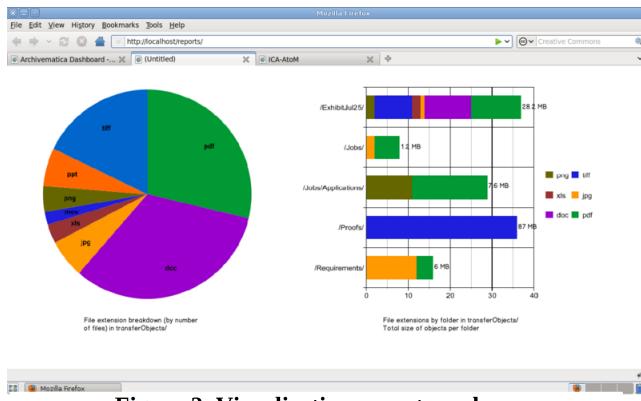


Figure 2. Visualization report mockup.

Such functions are being developed iteratively as part of the 2013 Archivematica 1.0 and subsequent releases. Should BitCurator or something else come along that can fulfill or expand on any of these functions, Archivematica's microservices architecture is such that the same requirements can be accomplished by these other tools with only minor changes to the code.

5. APPRAISAL

Originally intended for the long-term preservation of scientific data, OAIS does not address archival appraisal. To advise in the formation of appraisal requirements, the team consulted with the

InterPARES 3 Project [10] to conduct a gap analysis between OAIS and the InterPARES 1 Project's Chain of Preservation (COP) Model [11]. Review of the model, along with consultations with archivists about processing analogue records, revealed that appraisal occurs in a few different stages during archival processing. Archivists make an acquisition decision based on a preliminary appraisal, then reassess iteratively when they discover more about the records during accessioning actions and processing. Project partners including archivists and Archivematica developers built workflows around these different appraisal functions, which resulted in constructing three opportunities for appraisal in Archivematica: Selection for Acquisition, Selection for Submission and Selection for Preservation. The three appraisal opportunities as they were manifested in Archivematica 0.7.1 are discussed in detail in a recent Archivaria article [12], so the following is a brief summary of their functions and the associated Archivematica micro-services.

Selection for Acquisition occurs before records are accepted into an archives' custody for processing and preservation. Common practice in archives is to gather and review information about the records creator, the recordkeeping system(s) and the records to make an acquisition decision. For digital records, this includes learning as much as possible about the technological context of the records [8]. Because of limited access to originating technological environments for various reasons, it may become necessary for archives to acquire many more records than they might from an analogue body of records. Therefore, steps must be taken to ensure integrity of the records acquired while appraisal decisions are made over time.

Selection for Submission is the process of forming Submission Information Packages (SIPs) from acquired digital records or "transfers". In Archivematica, a transfer is any set of digital records acquired but not yet processed. Each SIP derives from one or more transfers. However, the SIP cannot be formed until the archivist has some information about the content of the transfer. For this reason, the transfer undergoes several micro-services first so that the archivist can review the results and assess how the received contents compare to the initial Selection for Acquisition expectations.

In the 0.8 alpha iteration of Archivematica, the archivist starts by adding a transfer to a specified folder in the file browser. The transfer begins processing in the Transfer tab of the web-based dashboard, where it is verified to be compliant for ingest in the system. Then, it is renamed with a transfer UUID and is assigned file UUIDs and checksums. If checksums already exist in the transfer, they are verified. A METS.xml file is added to the transfer, the transfer can be quarantined, and any packages are extracted. After a virus scan, prohibited characters are removed from filenames, formats identification process are run and metadata is characterized and extracted. All of the information generated from these micro-services allow the archivist to decide which parts of the transfer are archival materials ready for further processing.

In the 0.9 beta iteration of Archivematica, only one transfer can become one SIP, which is deprecated from the functionality of 0.8. In 0.8, one or more transfer(s) could become one or more SIP(s), but the arrangement was done in the file browser. The reason for this deprecation is so that the 1.0 release can move all the Selection for Acquisition functions to the web browser and

improve the tools to create SIPs as discussed in the Arrangement and Description section of this paper.

Selection for Preservation results in forming an Archival Information Package (AIP). A SIP is subjected to several micro-services, displayed in the Ingest tab, before the archivist has an opportunity to review the resulting AIP. Micro-services include verifying SIP compliance, renaming SIP with a SIP UUID, sanitizing file, directory and SIP name(s), checking integrity, copying metadata and logs from the transfer, and normalization. Once normalization and all other processing micro-services have run, the archivist can review the AIP contents and metadata in another browser window or download it to review using the file browser. At that point, they can either reject or accept the AIP and upload it into designated archival storage.

At every stage of appraisal, archivists may choose to destroy or deselect a record or set of records. Archivematica keeps logs of these changes by adding a text file listing excluded records to the logs directory in the transfer or SIP. This may even allow for richer and more transparent descriptive information about archival processing than is accomplished in analogue archives. It is important to note that the aforementioned steps are optional choices for the user. If the user has limited time or knows a great deal about the contents of a SIP, for instance, if the SIP is made up of described digitized videos, Archivematica can be configured to allow for automatic ingest.

In forthcoming releases, these appraisal processes will be incrementally moved to a web browser interface in the dashboard. Elastic Search indexing of the transfer and the AIP should also contribute to a richer, more informed selection process. Other development may include an automated process for “flagging” transfer content that may require further appraisal review based on a predefined set of indexing results.

6. PRESERVING AND PROVIDING ACCESS TO EMAIL

Several Archivematica project partners targeted email preservation as a priority in their digital archives planning. One pilot project involved acquiring a snapshot of the email account of a former university president. The account had been active for 10 years and no other email from the account had been sent to the university archives in electronic form in the past.

The university was using Zimbra Network Edition to send and receive email [13]. The Zimbra administrator's manual does not include information on how to export email from Zimbra for use in other email programs.[14] However, the university's IT department backs up the email accounts using a default directory structure specific to Zimbra, and was willing to deliver email to the Archives in the form of these backups. However, these backups are in a format which is intended to be used to restore email to Zimbra accounts, not to migrate the accounts' contents into other systems. Furthermore, documentation of its structure is somewhat limited. After analyzing the Zimbra backup and conducting research on email preservation standards and practices, the project team reached the conclusion that Zimbra email accounts need to be converted to a standard, well-documented, widely-used format that can be opened in a variety of open-source email programs or other tools such as web browsers.

Two formats which were explored as part of this project were Maildir and mbox [15]. Maildir is a text-based format which

stores each folder in an email account as a separate directory (inbox, sent items, subfolders etc) and each email as an individual text or .eml file [16]; attachments are included in the text files as base64 encoded ascii text. Mbox is a single large text file with attachments included as base64 content; each folder in an account is saved as a separate mbox file. Both formats can be imported into and rendered by numerous email programs, proprietary and open-source, and both can be converted into other formats using open-source tools and scripts. Although Maildir and mbox can be rendered in a variety of email programs, mbox has more potential as an access format because it is easier to develop tools to render it that are not necessarily email programs. For example, a program called Muse, developed by Stanford University [17], is designed to render mbox files using only a web browser. In addition, mbox is the source format for import into tools like the CERP email parser, which was developed by the Rockefeller Archive Center and the Smithsonian Institution Archives to convert email messages to hierarchically arranged XML files [18]. In essence, mbox is emerging as a de facto standard for which the digital curation community is beginning to build tools for rendering and manipulation. However, Maildir is preferable as a preservation format because it stores each message as a separate text file; thus any corruption to one or more text file would not cause an entire directory of messages to be lost, which is a risk with a format such as mbox.

The project team tested the use of a tool called OfflineImap [19] to back up a test Zimbra email account to Maildir and converted the Maildir backup to mbox using a freely available python script [20]. Following these preliminary tests, the Zimbra backup of the sample email account was restored to Zimbra and captured using OfflineImap. The resulting Maildir backup was converted to mbox files (Inbox, Sent and Eudora/out) which were imported into an open-source email program called Evolution. The total message count for each folder was found to be the same in Evolution as it had been in Zimbra (71, 2544 and 7628 messages, respectively), and randomly sampled emails were opened to ascertain that the conversion and import were successful. Sample emails from the Zimbra and Maildir backups were also compared to ensure that the significant characteristics of the Zimbra version were captured in the Maildir version [21].

A critical component of the University's email preservation strategy is management of access based on compliance with Freedom of Information and Protection of Privacy legislation. In any given user's account, some email messages must necessarily be excluded from public access based on the presence of personal information or other information which falls under exceptions to disclosure under the Act. The University's archivists and FOIPPA management personnel will need to be able to view email messages, flag those with restrictions, and provide public access to only those emails which are not restricted. Preliminary tests of Muse have shown it to be capable of importing mbox files, rendering the individual messages in a web browser, allowing tagging of restricted messages, and exporting the remainder in mbox format. We have noted that tagging one message as restricted automatically tags the same email message in other threads containing the same message.

Based on our analysis of pilot project email systems, email management practices, and preservation formats and conversion tools, we have summarized Archivematica requirements for acquiring, preserving and providing access to email. Ideally,

email is acquired, per account, in Maildir format, for the following reasons:

- The Maildir directory structure is well-documented and transparent;
- Maildir is widely used and can be created and rendered by a large number of software tools, both proprietary and open-source;
- OfflineIMAP is proving to be a useful tool for capturing email accounts in maildir format. Acting as an IMAP client, it can interact with a wide number of mail server programs, avoiding the need to add support for other mail server or email archive format conversions.
- The contents of a Maildir directory are plain text messages which can be read easily in any text editor (except for attachments);
- The text-based messages are based on an open and widely-used specification [22];
- Because each message is saved individually, accidental corruption or deletion of one or more messages would not result in the entire Maildir backup becoming unreadable (by comparison, corruption of a small amount of data in an mbox file could render the entire mbox file, with its multiple messages, unreadable);
- Maildir is easily converted to mbox for access purposes.

The archivists would submit the Maildir backup into Archivematica, where it would be retained as the preservation master in the AIP. Note that Maildir backups do not capture calendars or contact lists. However, University Archives staff have indicated that such records would probably not be considered archival. The attachments would be extracted and normalized to standard open formats for preservation purposes, with links between messages and their normalized attachments being managed through UIDs and/or filename. Attachments must be extracted and normalized because they pose a usability risk as base 64 ascii encoded text. They will always need to be rendered in a software program for human cognition of its content. In other words, even though the user may be able to open an email message in an email program he or she typically has to open the attachment separately using a software program that can render it.

For access, Archivematica will automatically generate a Dissemination Information Package (DIP) containing mbox files generated from the maildir preservation master. For an email account that consisted of an inbox with subfolders plus draft and sent items, the DIP would look something like this:

```
Inbox.mbox
Inbox.TravelCttee.mbox
Inbox.ExecCttee.mbox
Inbox.Workshops.mbox
Drafts.mbox
Sent.mbox
```

For most university and public repositories, provision of access must necessarily incorporate access and restriction management to comply with freedom of information, privacy and confidentiality requirements. The only known open-source tool that facilitates large-scale review and tagging of email account contents is Muse. More testing will be required to determine how usable and scalable the process of email tagging and exporting is with this tool. However, it should be noted that Muse is still in active development, and the Muse project team is interested in continuing to develop and refine the tool for use by libraries and archives. This bodes well for future feature development informed by Archivematica community members.

7. FORMAT POLICY REGISTRY - FPR

The Archivematica project team has recognized the need for a way to manage format conversion preservation plans, referred to by the project as format policies, which will change as formats and community standards evolve. A format policy indicates the actions, tools and settings to apply to a particular file format. The Format Policy Registry (FPR) will provide valuable online statistics about default format policy adoption as well as customizations amongst Archivematica users and will interface with other online registries (such as PRONOM and UDFR) to monitor and evaluate community-wide best practices. It will be hosted at archivematica.org/fpr.

An early prototype has been developed by Heather Bowden, then Carolina Digital Curation Doctoral Fellow at the School of Information and Library Science in the University of North Carolina at Chapel Hill (See Figure 3). A basic production version implementing these concepts will be included in upcoming releases. The FPR stores structured information about normalization format policies for preservation and access. These policies identify preferred preservation and access formats by media type. The choice of access formats is based on the ubiquity of viewers for the file format. Archivematica's preservation formats are all open standards; additionally, the choice of preservation format is based on community best practices, availability of open-source normalization tools, and an analysis of the significant characteristics for each media type. These default format policies can all be changed or enhanced by individual Archivematica implementers. Subscription to the FPR will allow the Archivematica project to notify users when new or updated preservation and access plans become available, allowing them to make better decisions about normalization and migration strategies for specific format types within their collections. It will also allow them to trigger migration processes as new tools and knowledge becomes available.

One of the other primary goals of the FPR is to aggregate empirical information about institutional format policies to better identify community best practices. The FPR will provide a practical, community-based approach to OAIS preservation and access planning, allowing the Archivematica community of users to monitor and evaluate formats policies as they are adopted, adapted and supplemented by real-world practitioners. The FPR APIs will be designed to share this information with the Archivematica user base as well with other interested communities and projects.

Extension	Normalization Description	Command Type	Command	Purpose
ac3	Transcoding to wav with ffmpg	bashScript	ffmpeg -i "%file.FullName%" -ac: ffprobe...	preservation
ac3	Transcoding to mp3 with ffmpg	bashScript	ffmpeg -i "%file.FullName%" -ac: ffprobe...	access
af	Transcoding to wav with ffmpg	bashScript	ffmpeg -i "%file.FullName%" -ac: ffprobe...	preservation
af	Transcoding to mp3 with ffmpg	bashScript	ffmpeg -i "%file.FullName%" -ac: ffprobe...	access

Figure 3 FPR format policies in early “Formatica” prototype. “Formatica” has since been renamed “FPR”.

8. CONCLUSION

Working with pilot project implementers, the Archivematica team has gathered requirements for managing a backlog of indexed digital acquisitions transfers, creating a SIP from a transfer or set of transfers, basic arrangement and description, preserving email, and receiving updates about new normalization paths via a format policy registry (FPR). After creating workflows that would account for real-world archival processing needs, these requirements have been added to our development roadmap for 0.9, 1.0 and subsequent Archivematica releases [23].

The Archivematica pilot project analysis and development described in this article are driven by practical demands from our early adopter community. The alpha release prototype testing sponsored by our contract clients and shared by a growing community of interested users from the archives and library professions and beyond has provided the opportunity to spearhead the ongoing evolution of digital preservation knowledge in the form of a software application that is filling a practical need for digital curators.

At the same time, the digital curation community is also evolving and maturing. New tools, concepts and approaches continue to emerge. The Archivematica technical architecture and project management philosophy are designed to take advantage of these advancements for the benefit of Archivematica users and the digital curation community at large.

The free and open-source, community-driven model provides the best avenue for institutions to pool their technology budgets and to attract external funding to continue to develop core application features as requirements evolve. This means the community pays only once to have features developed, either by in-house technical staff or by third-party contractors such as Artefactual Systems. The resulting analysis work and new software functionality can then be offered at no cost in perpetuity to the rest of the user community at-large in subsequent releases of the software. This stands in contrast to a development model driven by a commercial vendor, where institutions share their own expertise to painstakingly co-develop digital curation technology but then cannot share that technology with their colleagues or professional communities because of expensive and restrictive software licenses imposed by the vendor.

9. REFERENCES

- ISO 14721:2003, Space data and information transfer systems – Open archival information system – Reference model (2003).
- Artefactual Systems, Inc. and City of Vancouver, Requirements, <http://archivematica.org/wiki/index.php?title=Requirements> (accessed May 21, 2012).
- Artefactual Systems, Inc., Archivematica homepage, <http://archivematica.org> (accessed May 24, 2012).
- Archivematica significant characteristics evaluation, https://www.archivematica.org/wiki/Significant_characteristics (accessed August 19, 2012).
- Wikipedia definition of open standards, http://en.wikipedia.org/wiki/Open_standard (accessed August 17, 2012).
- Carolina Digital Repository Blog, “Announcing the Curator’s Workbench”, <http://www.lib.unc.edu/blogs/cdr/index.php/2010/12/01/announcing-the-curators-workbench/> (accessed May 21, 2012).
- BitCurator Tools for Digital Forensics Methods and Workflows in Real-World Collecting Institutions, <http://www.bitcurator.net/> (accessed May 21, 2012).
- Tika website, <http://tika.apache.org/> (accessed May 21, 2012).
- ElasticSearch website, <http://www.elasticsearch.org/> (accessed May 21, 2012).
- .InterPARES 3 Project, http://www.interpares.org/ip3/ip3_index.cfm (accessed May 21, 2012).
- .InterPARES 2 Project, Chain of Preservation (COP) Model, http://www.interpares.org/ip2/ip2_model_display.cfm?model=cop (accessed May 21, 2012).
- Courtney C. Mumma, Glenn Dingwall and Sue Bigelow, “A First Look at the Acquisition and Appraisal of the 2010 Olympic and Paralympic Winter Games Fonds: or, SELECT * FROM VANOC_Records AS Archives WHERE Value='true';” (Archivaria 72, Fall 2011) pgs. 93-122.
- Zimbra website, <http://www.zimbra.com/> (accessed May 21, 2012).
- Administration Guide to Zimbra, http://www.zimbra.com/docs/me/6.0.10/administration_guide/ (accessed May 24, 2012).
- Wikipedia articles describing maildir and mbox., http://en.wikipedia.org/wiki/Maildir_and_mbox. Note that this paper refers specifically to the .mbox extension, the standard Berkeley mbox implementation of this format. For another discussion of the role of mbox in email preservation, see Christopher J. Prom, “Preserving Email,” DPC Technology Watch Report 11-01 (December 2011), <http://dx.doi.org/10.7207/twr11-01>. (accessed May 23, 2012).

16. EML is a common email format encoded to the RFC 822 Internet Message Format standard (<http://tools.ietf.org/html/rfc822>) for individual emails. Messages in Maildir backups are encoded to this standard, although they lack the .eml file extension. For a discussion of the role in the eml format in email preservation, see Prom, “Preserving email”.
17. Muse website, <http://mobilisocial.stanford.edu/muse/> (accessed May 21, 2012).
18. CERP XML format, <http://siarchives.si.edu/cerp/parserdownload.htm>. The CERP XML format is designed to be a neutral, software-independent format for email preservation, but as yet there are no tools available to display the XML files as email messages that can easily be searched and navigated.
19. Offline Imap website, <http://offlineimap.org/> According to the documentation for this tool, it is possible to specify the folders to be captured, which would permit capturing folders designated specifically for archival retention. OfflineImap can also be run as a cron job, capturing email automatically at specified intervals. These features open up a number of possibilities for email archiving workflows.
20. Python script, md2mb.py, available from <https://gist.github.com/1709069>.
21. Significant characteristics analysis for maildir, http://www.archivematica.org/wiki/index.php?title=Zimbra_to_Maildir_using_OfflineImap for an example of the analysis of significant characteristics. (accessed May 24, 2012).
22. RFC # 822, Standard for the Format of ARPA Internet Text Messages, <http://tools.ietf.org/html/rfc822>.
23. Archivematica Development Roadmap, https://www.archivematica.org/wiki/Development_roadmap/ (accessed August 21, 2012).

Authenticity Management in Long Term Digital Preservation of Medical Records

Silvio Salza

Università degli studi di Roma "La Sapienza"
Dipartimento di Ingegneria informatica
via Ariosto 25, 00185 Roma, Italy
+39-06-77274-015
salza@dis.uniroma1.it

Maria Guercio

Università degli studi di Roma "La Sapienza"
Dipartimento di Storia dell'arte e spettacolo
piazzale Aldo Moro 5, 00185 Roma, Italy
+39-06-4967-002

maria.guercio@uniroma1.it

ABSTRACT

Managing authenticity is a crucial issue in the preservation of digital medical records, because of their legal value and of their relevance to the Scientific Community as experimental data. In order to assess the authenticity and the provenance of the records, one must be able to trace back, along the whole extent of their lifecycle since their creation, all the relevant events and transformations they have undergone and that may have affected their authenticity and provenance and collect the Preservation Description Information (PDI) as categorized by OAIS. This paper presents a model and a set of operational guidelines to collect and manage the authenticity evidence to properly document these transformations, that have been developed within the APARSEN project, a EU funded NoE, as an implementation of the InterPARES conceptual framework and of the CASPAR methodology. Moreover we discuss the implementation of the guidelines in a medical environment, the health care preservation repository in Vicenza Italy, where digital resources have a quite complex lifecycle including several changes of custody, aggregations and format migrations. The case study has proved the robustness of the methodology, which stands as a concrete proposal for a systematic and operational way to deal with the problem of authenticity management in complex environments.*

Categories and Subject Descriptors

H.3.2[Information storage and retrieval]: Information storage.

General Terms

Management, Documentation, Standardization, Legal Aspects.

Keywords

Authenticity, digital preservation, e-health, medical records.

1. INTRODUCTION

Authenticity plays a crucial role in the management and preservation of medical records. In most countries all the documentation related to the citizens' health, including of course digital files, has to be preserved for an indefinite period of time, some series potentially forever, and the continuing ability of assessing the authenticity and the provenance of the records is therefore an important

issue both for the legal value of data, to properly allocate the responsibilities, and for the scientific community that considers the results of medical tests and medical reports as important experimental data.

The problem of managing the authenticity of digital resources in this as well as in other environments has been addressed, as an important part of its activities, by the APARSEN project [1], a Network of Excellence funded by the EU (2011-2014) with the goal of overcoming the fragmentation of the research and of the development in the digital preservation area by bringing together major European players. The research activity we present here is the prosecution of the investigation carried out within previous international projects, notably the conceptual framework defined by InterPARES [6] and the methodology proposed by CASPAR [5]. More specifically, the APARSEN proposal [2] has stressed the need to take into account the whole *digital resource lifecycle* to model the preservation process, as defined by *ERMS* (*Electronic Records Management Systems*) recommendations, and has defined *operational guidelines* to:

- conveniently trace (for future verification) all the events and transformations the digital resource has undergone since its creation that may have affected its authenticity and provenance;
- collect and preserve for each of these events and transformations the appropriate evidence that would allow, at a later time, to make the assessment and, more precisely;
- develop a model of the digital resource lifecycle, which identifies the main events that impact on authenticity and provenance and investigate in detail, for each of them, the evidence that has to be gathered in order to conveniently document the history of the digital resource.

The model and the guidelines that we have proposed have been successfully put to test on experimental environments provided by the APARSEN project partners. These case studies, which are documented in a project deliverable [3], provided important feedback and have proved on the field the substantial robustness of the proposal.

This paper relates about a case study in the medical environment, the repository of the health care system in Vicenza (Italy), a rather complex case since along the DR lifecycle there are several changes of custody that involve, beside the preservation repository, several keeping systems, some of them geographically distri-

* Work partially supported by the European Community under the Information Society Technologies (IST) program of the 7th FP for RTD - project APARSEN, ref. 269977.

buted in the district. Moreover there are several types of DRs (diagnostic images, medical reports etc.), each one with a distinct workflow.

The case is also interesting because the repository must comply both with the international standards and with the rather complex Italian legislation on the creation, keeping and preservation of electronic records, and with additional specific rules for the keeping of medical records, based on the widespread use of digital signatures and certified timestamps. The implementation has been strongly oriented toward standardized solutions based on XML schemas) and a common dictionary based on PREMIS.

The paper is organized as follows. In Section 2 we present the Vicenza health care preservation system that has been the object of our study, and we also provide some details on the procedures mandated by the Italian legislation on long term preservation of digital records. Section 3 and 4 are devoted to present the digital resource lifecycle model and the authenticity management policy, as well as the operational guidelines that we propose to implement the model in specific environments and to guide the process of designing an effective authenticity management policy. In Section 5 these guidelines are applied to model the Vicenza health care system, and this leads to the formalization of the authenticity management policy and to the definition of the Authenticity Protocols. Finally, concluding remarks are given in Section 6.

2. THE VICENZA REPOSITORY

2.1 The preservation infrastructure

The preservation infrastructure of the public health care system unit ULSS6 in Vicenza is based on the system Scryba, implemented and distributed by the Italian company MEDAS Srl, that has been designed according to the basic principles of the OAIS reference model and with additional specific features intended to make it compliant with the Italian regulations on long term digital preservation. Scryba is a modular system based on a set of functionalities that can be configured to meet the specific requirements that arise in different environments. Up to now it has been deployed as the core element of several digital preservation repositories in Italian hospitals.

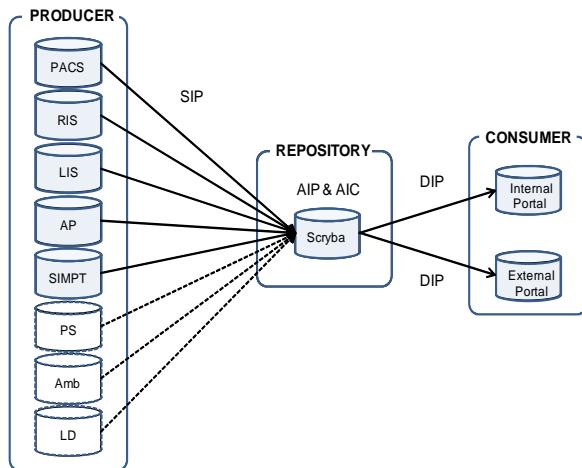


Figure 1. The preservation infrastructure.

In ULSS6-Vicenza the preservation infrastructure is interfaced with a variety of producers that deliver several different kinds of digital resources, mostly diagnostic images, test results and medical reports. The actual interface of the preservation system on the

producers' side is towards a set of departmental systems that collect the digital resources for peripheral devices and satellite systems, such as digital imaging devices, workstation attended by physicians etc.

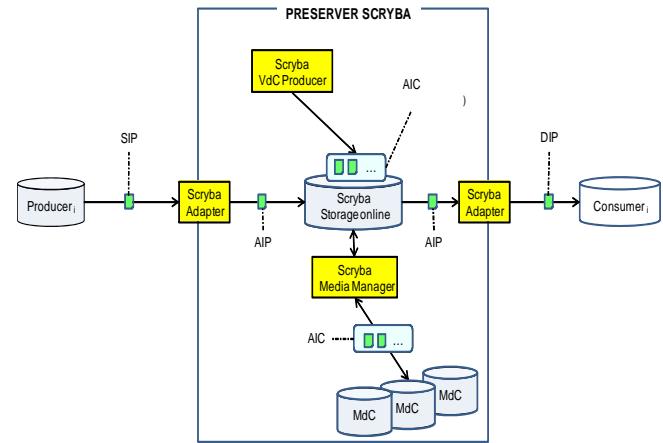


Figure 2. The Scryba preservation system.

The above mentioned departmental systems also act as short-term repositories and provide physicians and medical staff with immediate access to test results and reports. According to the Italian regulations, all medical records are delivered to the long-term preservation repository as soon as they are created and signed. Therefore, shortly after its creation and signature, each digital resource is preserved in two distinct copies, one in the departmental systems for consultation in the short period, and the other one in the *LTDP* (*Long Term digital Preservation*) repository as an official record.

The LTDP system can be accessed by consumers by means of two distinct interfaces:

- the internal portal which is used by physicians and medical staff, and allow authorized persons to get web access to the whole content of preserved digital resources;
- the external portal that provides citizens (or their authorized representatives) access to their own medical records.

Access to both interfaces requires strong authentication, according to the regulations on the privacy of medical records. An overview of the system is given in Figure 1 where the different kinds of producers are represented. Currently five different producers are supported, including diagnostic images in DICOM format (PACS) and medical reports of various kinds (RIS, LIS, AP). Support for additional producers is currently being implemented.

2.2 The Scryba preservation system

The Scryba system is based on the principles of the OAIS reference model and with additional specific features intended to make it compliant with the Italian regulations on long term digital preservation. The high level structure of the system is shown in Figure 2.

The system has a modular structure which is based on a core structure whose main functions are the management of the *AIPs* (*Archival Information Packages*), the related transformations (aggregation, format migration) and their secure storage. Additional modules, called *adapters*, are deployed to manage the communication with the external world, i.e. the *producers* on one side and the *consumers* on the other side.

Adapters are implemented on a base structure that can be customized to meet the specific requirements of different producers and consumers. Scryba Adapters work in several ways (DICOM protocol, HL7 msg, IHE XDS.b profile, or specific host oriented web-services) to match all host communication protocols. The management of the AIPs and their secure storage are compliant with the OAIS reference model, but strongly influenced by some peculiarities of the Italian national regulations. According to these regulations, the preservation process is based on collecting the digital resources to be preserved in large batches, named *Preservation Volume (PV)*, which are the actual object of the preservation process and must undergo a well-defined formal procedure that includes digital signature, certified time stamping of the PV as well as periodical controls and possibly the generation of new copies on different storage medias.

The Italian regulations require also to produce a given number of *BCs* (*Backup Copies*) for every PV and to store them in different locations according to a predefined and formally stated schema.

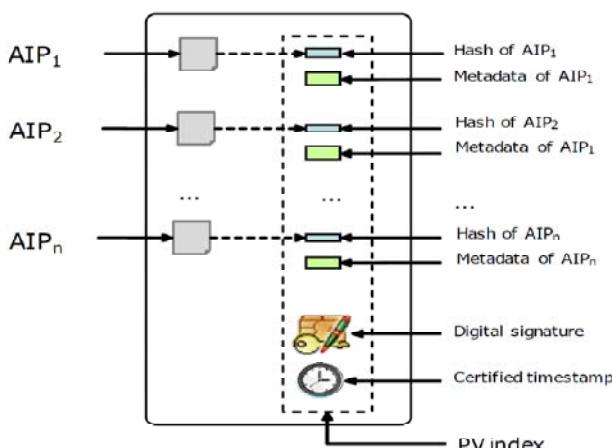


Figure 3. Structure of a Preservation Volume.

The structure of a preservation volume is shown in Figure 3. It contains all the aggregated digital resources plus an additional file, the *Preservation Volume index* (PV index), which is compliant to UNI SInCRO, a national metadata standard, digitally signed by person officially in charge of the preservation process (in Italian *Responsabile della Conservazione*) and marked with a temporal timestamp. The PV index is an XML file which contains:

- a hash file for each AIP in the PV;
- a set of metadata for each AIP in the PV;
- the digital signature;
- the certified timestamp.

In order to comply both with the OAIS model and the Italian regulations, the SIPs are ingested as soon as they are delivered to the Scryba system, and an AIP is generated for each SIP, i.e. for every individual study or medical report, and enters immediately the preservation process. On the other hand, a set of AIPs from each producer is periodically aggregated to generate an AIC (Archival Information Collections), an OAIS kind of Information Package that well corresponds to the PV (Preservation Volume) the Italian regulations ask for. In the Scryba system any given PV must contain digital resources of a single type and PVs are closed according to a double criteria:

- time: a PV must be closed before a maximum time since its opening elapses (currently 24 hours);

- size: a PV cannot exceed a maximum size. (currently 1GB).

We shall point out that the aggregation of several digital resources in a single preservation volume only depends on the national regulations and it is not performed to comply with OAIS.

3. THE AUTHENTICITY MODEL

3.1 The digital resource lifecycle

The main principle behind the authenticity management methodology that has been developed within the APARSEN project is that, in order to properly assess the authenticity of a Digital Resource (DR), we must be able to collect the information relevant for preservation and trace back, along the whole extent of its lifecycle since its creation, all the transformations the DR has undergone and that may have affected its authenticity and provenance. With specific reference to the transformations crucial for LTDP, for each of these transformations one needs to collect and preserve the appropriate evidence that would allow, at a later time, to make the assessment, and that we shall call therefore *authenticity evidence*.

Under quite general assumptions, we may consider the DR lifecycle as divided in two phases:

- **Pre-ingest phase.** This phase begins when the DR is delivered for the first time to a keeping system and goes on until the DR is submitted to a *Long Time Digital Preservation (LTDP) system*. During the pre-ingest phase, the DR may be transferred between several keeping systems and may undergo several transformations, and is finally transferred to the LTDP system.
- **LTDP phase.** This phase begins when the DR is ingested by a LTDP system and goes on as long as the DR is preserved. As for the pre-ingest, also during the LTDP phase the DR may undergo several transformations, notably format migrations, aggregations etc. Moreover it may get moved from a LTDP system to another one.

The pre-ingest phase has been introduced as a separate phase from the ingest to represent the part of the lifecycle that occurs before the delivery to the DR of a LTDP system. Collecting evidence for all the transformations the DR undergoes during this phase is of the utmost importance to assess its authenticity.

Each transformation a DR undergoes during its lifecycle is connected to an *event*, which occurs under the responsibility of one or more people, whom we shall call *agents*. A transformation may involve one or several DRs and one or several agents, and produces as a result a set of DRs, possibly new versions of the ones that were the object of the transformations.

Unfortunately, the variety of events that may occur during the DR lifecycle is very large and depends, at least in part, from the specific environment. Nevertheless, it is possible to consider at least a minimal *core set of events*, that includes the most important ones, as well as the ones which are likely to occur in most of the environments in which DRs are produced and managed. The core set, is briefly discussed in the following subsections, and may be considered as a preliminary step towards interoperability in the exchange of authenticity evidence among different keeping and preservation systems.

In our investigation we have considered a reasonable variety of environments, notably natural science data, health care data, social science data and administrative data repositories. As a result of our analysis, we have proposed the core set of events that we briefly outline. For a more complete description one should refer directly to the APARSEN project documentation [2].

3.2 Pre-ingest phase

During its stay in the keeping system the DR may undergo a series of transformations that may affect both its content and the descriptive information associated to it. For instance the DR may go through format migrations (even before it enters the LTDP custody), or it may get integrations of its content and/or of its metadata, or it may eventually be aggregated with other DRs to form a new DR. Moreover, before getting to LTDP, the DR may be transferred, one or several times, between different keeping systems.

The pre-ingest phase includes also the submission of the DR to the preservation repository. The content and the structure of the SIP (Submission Information Package) through which the DR is delivered must comply with a submission agreement established between the system where the DR was kept (i.e. the Producer in the OAIS reference model) and the LTDP system (the OAIS).

In the model, the core set for the pre-ingest phase comprises the following events:

- **CAPTURE**: the DR is delivered by its author to a keeping system;
- **INTEGRATE**: new information is added to a DR already stored in the keeping system;
- **AGGREGATE**: several DR, already stored in the keeping system, are aggregated to form a new DR;
- **DELETE**: a DR, stored in the keeping system is deleted, after its preservation time has expired, according to a stated policy;
- **MIGRATE**: one or several components of the DR are converted to a new format;
- **TRANSFER**: a DR is transferred between two keeping systems;
- **SUBMIT**: a DR is delivered by the keeping system where it is stored (producer) to a LTDP system.

3.3 LTDP phase

This phase begins when the DR is delivered to a LTDP system and goes on as long as the DR is preserved. During this phase, the DR may undergo several kinds of transformations, that range from format migrations to changes of physical support, to transfers between different preservation systems.

According to the OAIS reference model [4], many activities are carried out in connection with each of these events, but we restricted our attention to the sole aspects related to authenticity and provenance of the DR and to the information (authenticity evidence) that has to be gathered and preserved in the PDI (Preservation Description Information), and more specifically in the Provenance, Context and Fixity components.

Analyzing this phase many possibilities have to be considered, as for instance transfer between LTDP systems, which is quite likely to happen in the long run, and changes in the structure of the preserved DRs (integration, aggregation etc.), that routinely happen in the health care sector, since records must enter preservation as soon they are created and still there may be later the need to introduce corrections.

The resulting set of events is then:

- **LTDP-INGEST**: a DR delivered from a producer is ingested by the LTDP system and stored as an AIP;
- **LTDP-AGGREGATE**: one or several DRs stored in different AIPs, are aggregated in a single AIC;
- **LTDP-EXTRACT**: one or several DRs which are extracted from an AIC to form individual AIPs;

- **LTDP-INTEGRATE**: new information is added to a DR already stored in the LTDP system;
- **LTDP-MIGRATE**: one or several components of a DR are converted to a new format;
- **LTDP-DELETE**: one or several DR, preserved in the LTDP system and stored as part of an AIP are deleted, after their stated preservation time has expired;
- **LTDP-TRANSFER**: a DR stored in a LTDP system is transferred to another LTDP system.

3.4 Event templates

When giving the guidelines that should be followed to ensure interoperability on authenticity among keeping and LTDP systems, beside providing a precise definition of the event, the crucial point is to specify which controls should be performed, which evidence should be collected and how it should be structured.

In the model each event of the core set is represented according to a uniform schema, by providing an *event template*:

- the *agent*, i.e. the person(s) under whose responsibility the transformation occurs;
- the *input*, i.e. the preexisting DR(s) that are the object of the transformation, if any;
- the *output*, i.e. the new DR(s) that are the result of the transformation (possibly new versions of input DR(s));
- the *controls* that must be performed when the event occurs on the authenticity and provenance of the input DR(s) and to assess properties of the output DR(s) that are the results of the transformation connected to the event.
- the *Authenticity Evidence Record (AER)*, i.e. the information that must be gathered in connection with the event to support the tracking of its authenticity and provenance.

An event template is therefore a sort of checklist, enumerating all the controls that should be performed and all the authenticity evidence that should be gathered and preserved in order to guarantee an accurate management of the DR authenticity through its lifecycle.

Event templates have been defined in the model under very general assumptions, and therefore have been developed into very comprehensive checklists. That means that in a given specific environment only part of the controls may actually need to be performed and only part of the authenticity evidence that is listed in the AER may actually need to be gathered.

Therefore, the model and the templates should be considered as a very general and detailed reference, that needs accurate customization in each specific environment to get to the definition of an adequate authenticity management policy, a problem that will be addressed in Section 4.

3.5 Authenticity Evidence Records

A crucial part of the event template is the definition of the *Authenticity Evidence Records (AER)*. An AER is specified as a sequence of *Authenticity Evidence Items (AEIs)*, i.e. of the elementary items of information that should be gathered and preserved to document the authenticity and the provenance of the DR.

As the DR progresses along its lifecycle through a sequence of events, an incremental sequence of AERs, that we shall call *Authenticity Evidence History (AEH)*, is collected by the systems where the DR is kept or preserved, and strictly associated to it.

From a practical point of view, an authenticity evidence record is a structured set of information, according to our proposal an XML

file of predefined structure, which is strictly related to a given event. At any given stage of its lifecycle a DR brings with it, as part of its metadata, a (temporally) *ordered sequence* of such records, to document all the transformations the DR has undergone and to allow to assess its authenticity and provenance.

Authenticity evidence will follow the DR when it is transferred between different systems, and will accompany it along all its lifecycle. Thus, to ensure interoperability, it is necessary to standardize the way the authenticity evidence is collected and structured. To this purpose existing standards should be accurately considered, as for instance the Open Provenance Model (<http://openprovenance.org>).

4. THE OPERATIONAL GUIDELINES

Aim of this section is to present the procedure, i.e. the sequence of steps, that should be followed, when dealing with the problem of setting up or improving an LTDP repository in a given specific environment, to get to the definition of an adequate *authenticity management policy*, that is to formalize the rules according to which authenticity evidence should be collected, managed and preserved along the digital resource lifecycle.

4.1 Role of the Designated Community

The concept of *Designated Community (DC)* ("an identified group of potential Consumers who should be able to understand a particular set of information") is central to the OAIS reference model according to which "the primary goal of an OAIS is to preserve information for a designated community over an indefinite period of time". Therefore, as a first step, one should understand what authenticity means to the DC, that is:

- for which purpose and to which extent is the DC interested in being able to assess the authenticity and the provenance of the DRs that are preserved by the OAIS?
- what kind of evidence is considered by the DC as sufficient to make the assessment?

When dealing with an existing LTDP repository, that is analyzed to assess the adequacy of the current practices or to suggest improvements, the starting point should be understanding what kind of authenticity evidence is currently preserved and investigating if the DC actually deems it as sufficient for its purposes.

Altogether the result of this preliminary step is to set up a reference context in order to take appropriate decisions in the following steps of our procedure, i.e. when identifying the lifecycle events to be taken into account and the specific authenticity evidence to be gathered in connection with them.

4.2 Identifying the relevant lifecycle events

The next step is to analyze the workflow of the DRs that are to be preserved in the repository, from their creation on, to identify the lifecycle events that are relevant to the management of the authenticity. When, as it was in the Vicenza case, several DR types and several workflows are identified, the analysis is to be repeated for each workflow.

Once the relevant lifecycle events have been identified, they must be compared and fitted into the *core set events* that we have discussed in Sections 3.2 and 3.3 and that provide a reference and a template on the way authenticity evidence should be gathered and managed. According to our case study experience the core set that we have proposed has proved to be a robust choice, in the

sense that all the relevant events we have identified could fit well in one of the core set events. However, it is still possible that in a given environment additional events may need to be considered that are specific to that environment.

Then, for each lifecycle event we have identified as relevant, the corresponding event templates should be considered to identify responsibilities and to understand which authenticity evidence should be gathered and which controls should be performed.

As we already pointed out, the templates are quite comprehensive. Therefore, it is often found that part of the authenticity evidence that the templates mandate to collect is not actually collected in the current practices. This does not necessarily mean that the current practices are inadequate: one should instead carefully consider every single missing item of evidence, taking into account the specific needs of the designated community and other details, as for instance the systems involved and their ownership.

For instance, criteria for deciding if an authenticity evidence item should *not necessarily* be recommended as part of the AER could be:

- the item is intended to document a control that is actually performed but not recorded in the AER by a system under the ownership of an organization which is trusted by the DC;
- the item is intended to prove that the integrity of the item has not been affected by the transfer between two systems that are under the ownership of the same organization which is trusted by the DC;
- the item relates to some provenance information which is of no interest to the designated community.

Anyway, besides a few general criteria as above, it is difficult, probably impossible, to give an exhaustive list of specific criteria for deciding whether a given authenticity evidence item should be recommended or not, mostly due to the variety of situations and the complexity of systems.

4.3 Defining the policy and the authenticity evidence records

As a result of the analysis performed in the previous step one should be able to reach, for any given authenticity evidence item in the template of a given lifecycle event, one of the following conclusions:

- a) the evidence item is currently collected and preserved and must be part of the AER;
- b) the evidence item is not currently collected and preserved, but this information is not necessary according to the definition of authenticity that is accepted by the DC;
- c) the evidence item is not currently collected and preserved, but it is not possible to prove that this information is not necessary, and it must therefore become part of the AER.

In all three cases the conclusions should be explicitly and clearly documented. In case c) an improvement of the current practices should be recommended and the information to be collected should be clearly specified, along with the procedure to collect it.

The result of all the above actions is the definition of the *authenticity management policy* that should be adopted by a given LTDP repository to comply with the guidelines we propose and satisfy the needs of its DC. This is made up of the following components:

- i. a general statement about the meaning of authenticity to the DC, accompanied by a clear delimitation of the DC and by the

- explanation of how the opinion of the DC was actually gathered;
- ii. the specification of the lifecycle, and more precisely of the events in the lifecycle that have been identified as relevant to the management of authenticity;
 - iii. for every relevant event in the lifecycle the definition of the controls corresponding to that event that must be performed and of the AER, together with the specification of the procedures that should be followed to collect it.

4.4 Formalizing authenticity protocols

The next step in implementing the authenticity management policy is the formal definition of the controls that must be performed in connection with each event and of the procedures that must be followed to collect the AER. To this purpose, we propose an implementation strategy which is based on the concept of *Authenticity Protocol (AP)* that has been defined within the CASPAR project [5] as the specification of the procedure that must be followed to assess the authenticity of specific type of DR.

In our methodology the AP becomes the procedure that must be followed in connection with a given lifecycle event to perform the controls and to collect the AER as specified by the authenticity management policy. Accordingly, the execution of the AP corresponding to a give lifecycle event generates the AER that the authenticity management policy mandates to collect in correspondence to that event.

In the formal definition an AP is characterized by:

- *DR type*: the type of digital resource
- *Event type*: the lifecycle event to which the AP corresponds
- *Agent*: the person under whose responsibility the protocol is executed
- *AER*: the AER that is generated by the execution of the AP
- *AS sequence*: the sequence of authenticity steps (AS) that must be performed

In turn, every AS in the AP consists in a set of elementary actions meant to perform a specific control and/or to collect one or more authenticity evidence items, and is characterized by:

- *Controls*: the set of controls that must be performed
- *Input*: the items from the content of the processed DR and its AEH on which the AS operates
- *Output*: the set of authenticity evidence items generated by the execution of the AS
- *Actions*: a set of additional actions that are (possibly) performed as a result of the controls

Defining the APs is therefore a long and repetitive process, though a rather systematic one once the procedure is established.

5. THE VICENZA CASE STUDY

5.1 Modeling the DR lifecycle

As part of the APARSEN project activities, the Vicenza health care system preservation repository, that we have discussed in Section 2, has been selected as one of the test environments for the implementation of the authenticity model that we have presented in the previous sections. In this case study the APARSEN authenticity management guidelines have been applied to their full extent, i.e. from the preliminary analysis to the formal definition of the authenticity management policy, that is to the specification of the APs. Referring to the guidelines has provided valuable help, both in pointing out any weakness in the current practices and in providing a reasonable way to fix the problems.

In this section we shall discuss in some detail the management of medical reports, one of the several DR types which are managed by the Scryba preservation system. Further details can be found in the project documentation [3].

Medical reports are written by physicians to interpret and comment studies of diagnostic images, to which they are connected through the accession number. Reports are written using a specific *Radiology Information System (RIS)* application which is run on local systems, and are digitally signed by the physicians who write them. The digital signature process, which is directly managed by the RIS application, follows the Italian regulations and is based on the digital certificate of the physician which is held in his own smart-card or in a *HSM (High Security Module)* device for remote signature. As soon as they are completed reports are stored in a central archive managed by a centralized RIS.

According to the Italian regulations, digitally signed reports are in pkcs#7 format, a cryptographic envelope that contains:

- the report;
- the digital certificate of the physician;
- a hash file of the report encrypted with the private key of the physician.

The above information is of crucial importance to assess the authenticity and provenance of the report.

Reports are submitted by the RIS system to the preservation system almost as soon as they are completed (an upload procedure is run every 5 minutes). A SIP is generated for every single report, which is made up of two components:

- the pkcs#7 (i.e. report + certificate + signature);
- a XML metadata file.

Metadata include:

- DICOM identifier of the study to which the report refers
- Version ID (several versions of the report may be submitted and must be treated as different documents)
- Patient ID
- Patient Name
- Patient birth date
- Patient gender
- Date of the exam

As soon as a SIP is accepted by the repository, a unique identifier (ID-DOC-Scryba) is assigned to the digital resource and a confirmation message is sent to the RIS. Then a set of controls are performed during the ingestion process:

- *Unicity check*: a check is performed to check in the repository database that the given report with the same version number and the same hash is not already in the repository.
- *Provenance check*: the digital certificate contained in the pkcs#7 file is checked against the information downloaded from the certification authority (original certificate and revocation list). This check guarantees the identity of the physician who has signed the report, and hence its provenance.
- *Fixity check*: the digital signature is decrypted and the resulting hash is compared against the hash of the report component of the pkcs#7 file. This check guarantees the integrity of the report.

Moreover a certified timestamp of the report is generated. This guarantees the existence and the content of the report at the time the timestamp is generated. In Italy the timestamp has a legal validity of 20 years.

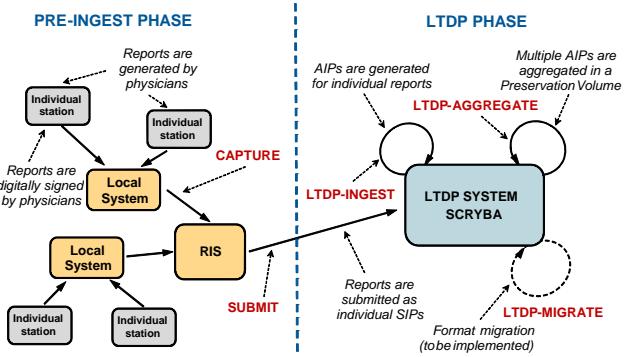


Figure 4. RIS lifecycle model.

The RIS workflow lifecycle can be conveniently modeled according to the APARSEN guidelines, and all events which are relevant to the management of authenticity, namely changes of custody and transformations of the digital resources, prove to fit well in the core set events. The resulting lifecycle model is shown in Figure 4. In the picture the two lifecycle phases, the *pre-ingestion phase* and *LTDP phase* are clearly identified, as well as the five events that we consider relevant for the management of authenticity: CAPTURE, SUBMIT, LTDP-INGEST, LTDP-AGGREGATE and LTDP-MIGRATE.

5.2 Defining the policy

The next step is, according to the guidelines, for each lifecycle event, to compare the controls and the authenticity evidence recommended by the event templates with the current practices in the repository (see Section 4.2). This analysis has pointed out that some of the controls are currently missing and that some of the authenticity evidence is not gathered. It is therefore necessary to carefully investigate if there is a solid justification for this.

It actually turns out that the lack of part of the authenticity evidence items that are recommended by the templates is the result of the following assumptions by the repository management, which are in turn based on a general notion of trust:

- all transfers among systems are carried on private lines that are under the ownership of a single administration (the Vicenza Public health care system), and are managed with adequate security provisions;
- access to the systems is given only to registered users, and a proper rights management policy is enforced;
- reports, after they are generated, get to the preservation repository in a very short time, therefore threats to their integrity can be considered as negligible.

These assumptions are indeed quite reasonable, and altogether we may rate the current practices in handling this event as acceptable, as long as one makes clear that:

- no controls are performed and no evidence is documented when the DRs are transferred between systems in the pre-ingestion phase;
- the integrity of data and metadata strictly depends on trusting the whole infrastructure under the ownership of the Vicenza Public health care system.

These issues and the related threats should be carefully discussed with the Designated Community, who should clearly confirm its understanding and its consensus. A preliminary analysis shows that the main (and perhaps the only) concern of the DC is the

compliance with the national regulations on LTDP, which actually can be proved.

Nevertheless one should consider that the DRs we are dealing with may become evidence in court cases about forgery or loss of data, and therefore it may be necessary to prove that their integrity has been maintained in a more substantial way. It can be argued that substantial evidence in proving the integrity could come from system logs and from the rights management policies, but this raises the further question of how long this information is maintained and how it is preserved.

Therefore we would like to suggest that some additional authenticity evidence should be preserved, for instance, for every transfer of the digital resource, a record of the time of the transfer and the identification of the source and destination system administrators.

5.3 Implementing authenticity protocols

To implement the authenticity management policy it is necessary to define the authenticity protocols for all the events in the lifecycle model. In this section we give, as an example, the authenticity protocol for the event INGEST. According to our methodology (see Section 4.4) the protocol consists in the specification of all controls and actions that must be performed during the ingestion to check the authenticity and the provenance of the DR and to generate the Authenticity Evidence Record (AER), which comprises the following *Authenticity Evidence Items (AEIs)*:

- AEI-1. *Event type*: ingest
- AEI-2. *Original identifier*: identifier from the report metadata.
- AEI-3. *New identifier in the LTDP system*: ID-DOC generated by Scryba
- AEI-4. *Context information*: DICOM identifier of the study to which the report refers.
- AEI-5. *Date and time the ingestion has been completed*: from the certified timestamp
- AEI-6. *Identification and authentication data of the LTDP system administrator*: generated by Scryba
- AEI-7. *Assessment on the authenticity and provenance*: outcome of controls on the digital signature
- AEI-8. *Digest of the AIP*: from the certified timestamp.

As discussed in Section 4.4, the protocol consists in a general specification (DR type, event type, agent etc.) and in a sequence of AS, each meant to perform a specific control and/or to collect one or more authenticity evidence items:

- DR type: RIS - Digitally signed medical reports
- Event type: LTDP-INGEST
- Agent: administrator of the Scryba system
- AER: as defined above
- AS sequence: steps from AS-1 to AS-12

The individual authenticity steps are detailed as follows:

STEP AS-1 - CHECK PROVENANCE

- AS-1.1: get the digital signature certificate from the pkcs#7 file
- AS-1.2: get the original digital certificate from the Certification Authority
- AS-1.3: check the certificate in the pkcs#7 file against the original certificate
- AS-1.4: check the expiration date in the digital certificate against the current date
- AS-1.5: get the revocation list from the Certification Authority and check it

- **AS-1.6:** if any of the checks in **AS-1.3**, **AS-1.4** and **AS-1.5** fails then abort ingestion

STEP AS-2 - CHECK INTEGRITY

- **AS-2.1:** generate the hash file of the report component in the pkcs#7
- **AS-2.2:** decrypt the digital signature in the pkcs#7 file by using the public key
- **AS-2.3:** compare the two hash files generated in steps AS-2-1 and AS-2.2
- **AS-2.4:** if the check in **AS-2.3** fails then abort ingestion

STEP AS-3 - CHECK CONTEXT

- **AS-3.1:** extract the identifier of the study to which the report refers from **AER RIS-CAPTURE**
- **AS-3.2:** check the Scryba DB to verify that a study exists with identifier generated in step **AS-3.1**
- **AS-3.3:** if the check in **AS-3.2** fails then abort ingestion

STEP AS-4 - GENERATE INTERNAL IDENTIFIER

- **AS-4.1:** generate an internal unique identifier that identifies the DR in the repository

STEP AS-5 - GENERATE TIMESTAMP

- **AS-5.1:** generate a hash file of the content information of the AIP
- **AS-5.2:** send the hash file generated in **AS-5.1** to the Certification Authority to get a certified timestamp;

STEP AS-6 - GENERATE AEI: Original Identifier

- **AS-6.1:** generate AEI-2. *Original identifier* which is given the value extracted in **AS-4.1**.

STEP AS-7 - GENERATE AEI: Internal Identifier

- **AS-7.1:** generate an internal unique identifier for the DR in the Scryba system
- **AS-7.2:** generate AEI-3. *New identifier in the LTDP system* which is given the value generated in **AS-7.1**

STEP AS-8 - GENERATE AEI: Context Information

- **AS-8.1:** generate AEI-4. *Context information* which is given the value extracted in **AS-3.1**.

STEP AS-9 - GENERATE AEI: Date And Time

- **AS-9.1:** extract date and time from the certified timestamp
- **AS-9.2:** generate AEI-5. *Date and time the ingestion has been completed* which is given the value extracted in **AS-9.1**.

STEP AS-10 - GENERATE AEI: Administrator Data

- **AS-10.1:** generate AEI-6. *Administrator data* with the Scryba system administrator data

STEP AS-11 - GENERATE AEI: Assessment on Authenticity and Provenance

- **AS-11.1:** generate AEI-7. *Assessment on authenticity and provenance* which documents the outcome of the checks performed in **AS-1** to **AS-4**

STEP AS-12 - GENERATE AEI: DIGEST OF THE AIP

- **AS-12.1:** generate AEI-8, *Digest of the AIP* which is given the value of the hash file generated in **AS-6.1**.

6. CONCLUDING REMARKS

In this paper we have presented the model we propose for the management of the authenticity of the digital resources through their lifecycle, including the LTDP phase, and the operational guidelines for its deployment and the definition of the authenticity management policy in a specific environment. Moreover we have reported a case study, a repository of medical records, in which the methodology has been successfully tested.

The case study has been a quite interesting and fruitful experience, both for our team, which was concerned with the testing of the methodology and for the management of the repository which was interested in assessing the current practices and in devising possible improvements. The specific environment was indeed well suited for the purpose in several ways:

- the designated community shows a clear interest (and a strong commitment) in the problem of properly managing authenticity and provenance of DRs;
- the repository manages a variety of DRs and with quite a reasonable lifecycle complexity (changes of custody and transformations of the DRs);
- the repository has to comply with the quite demanding and detailed Italian rules on LTDP and the keeping of medical records, which mandate authentication of the records through digital signatures and certified time stamping, and consequently provide crucial evidence on the integrity and provenance of the records.

The model has proved to be robust enough and allowed to conveniently accommodate all the transformations and the changes of custody in the workflow. On the other hand, the templates provided by the model for the authenticity evidence records have been a comprehensive checklist to verify which authenticity evidence was actually gathered in the current practices of the repository, and to understand what information was missing and which improvements should be possibly suggested.

Another positive outcome of the case study was to confirm the flexibility of the approach that we propose, that is the ability to guide the definition of an authenticity management policy tailored to the needs of the specific environment. This is indeed a crucial issue, since different communities may have different needs and may attach to the concept of authenticity a different meaning and a different value. The balance between cost and effectiveness may therefore have quite different points of equilibrium.

7. REFERENCES

- [1] APARSEN Project – Alliance Permanent Access to the Records of Science in European Network (2011-2014), <http://www.alliancepermanentaccess.org/>
- [2] APARSEN Project: D24.1. Report on Authenticity and Plan for Interoperable Authenticity Evaluation System (2012) http://aparsen.digitalpreservation.eu/pub/Main/ApanDeliverables/APARSEN-DEL-D24_1-01-2_3.pdf
- [3] APARSEN Project: D24.2. Implementation and testing of an Authenticity Protocol on a Specific Domain. (2012) http://aparsen.digitalpreservation.eu/pub/Main/ApanDeliverables/APARSEN-DEL-D24_2-01-2_2.pdf
- [4] CCSDS: Reference Model for an Archival Information System – OAIS. Draft Recommended Standard, 650.0-P-1.1 (Pink Book), Issue 1.1 (2009), <http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/CCSDSAgency.aspx>
- [5] Factor M., Guercio M., et al.: Authenticity and Provenance in Long Term Digital Preservation: Modeling and Implementation in Preservation Aware Storage. TaPP '09 (2009) http://www.usenix.org/event/tapp09/tech/full_papers/factor_factor.pdf
- [6] InterPARES Projet: Requirements for Assessing and Maintaining the Authenticity of Electronic Records (2002), http://www.interpares.org/book/interpares_book_k_app02.pdf

Future-Proof Preservation of Complex Software Environments

Klaus Rechert, Dirk von Suchodoletz and Isgandar Valizada

Department of Computer Science
University of Freiburg, Germany

ABSTRACT

Emulation evolves into a mature digital preservation strategy providing authentic access to a wide range of digital objects using their original creation environments. In contrast to migration, an emulation approach requires a number of additional components, namely the full software-stack required to render a digital object, and its configuration. Thus, independent of which emulator is chosen, contextual information of the original computer environment is always needed.

To overcome this knowledge gap, a formalization process is required to identify the actual building blocks for an authentic rendering environment of a given object. While the information gathering workflow relies heavily on user knowledge and manual interaction during ingest, the workflow is coupled with a feedback loop so that both a complete emulation environment and preservation of desired properties for later access are ensured.

1. INTRODUCTION

In most cases the best way to render a certain digital object is using its creating applications, since these cover most of the object's significant properties and hence, provide an authentic and possibly an interactive user experience. Therefore, emulation is a key strategy to provide a digital object's native environment and thus to maintain its original "look" and "feel" [7]. In some cases the existence of access alternatives, e.g. format migration, is not guaranteed due to the proprietary nature of the object's file formats, or even impossible due to the complex structure of the object (e.g. digital art, computer games, etc.).

Recreating software environments using emulation requires detailed knowledge about the objects' dependencies (e.g. operating systems, libraries, applications). Viewpaths (VP) [6] represent an ordered list of such dependencies for a given object, defining an order for the sequence in which these dependencies are required. Hence, in combination with a comprehensive and well-managed software archive any ancient computer environment could be rebuilt. Nevertheless this topic is still largely neglected by practitioners and research communities in the digital preservation domain.

If a digital object becomes subject to digital preservation, a defined workflow is required to support the preservation process of the object's original context i.e. rendering environment. The workflow makes use of the user's knowledge to identify necessary components of the object's rendering environment, to the effect that the rendering environment is complete and there are no dependency conflicts, at least

for the chosen configuration and the digital object's contextual environment. For current computer environments and applications plenty of user knowledge is available. Thus, the project's proposed workflows focus on users "owning" a system setup e.g. for performing business of scientific processes. More specifically, owners of today's digital objects have good knowledge of the object's properties, their desired functions and utility, at least to extent of the objects' original purpose. Furthermore, preserving the knowledge of installation, configuration, and usage of software components ensures the recreation process of past system environments. By providing a preview of the emulated and recreated environment during ingest the user is able to test if the chosen setup meets the desired rendering quality and functionality. Figure 1 shows the proposed workflow in an abstract way.

2. RELATED WORK

In order to preserve a digital object's rendering environment, any dependencies from interactive applications to operating system and hardware components need to be identified. A widely adopted method which is integrated as a service in many digital repositories and institutional archives is the file type database PRONOM [2]. But identifying files and linking applications to them is only the first step. Several tools were proposed to resolve software dependencies from platform specific object-code binaries. E.g. DROID¹ makes use of "file-magic" fingerprints in combination with a database, others make use of system library resolving mechanisms [4]. While these tools and techniques provide useful information and hints to the users, they do not guarantee the generation of a suitable rendering environment, for instance, regarding completeness, quality and conflicting dependencies. In case of database dependent tools, appropriate data for a specific digital object is required.

A significant challenge when dealing with outdated software packages is the diminishing knowledge of how to handle the installation and configuration processes properly. One method to leverage the effort and archive the required knowledge is to automate the different installation steps for each relevant package. A viable approach is illustrated by Woods and Brown, who describe a software designed to minimize dependency on this knowledge by offering automated configuration and execution within virtualized environments [10]. This group demonstrated how to deploy automation scripts, i.e. GUI automation in order to install applications

¹DROID Project, <http://droid.sourceforge.net/>, (5/28/2012).

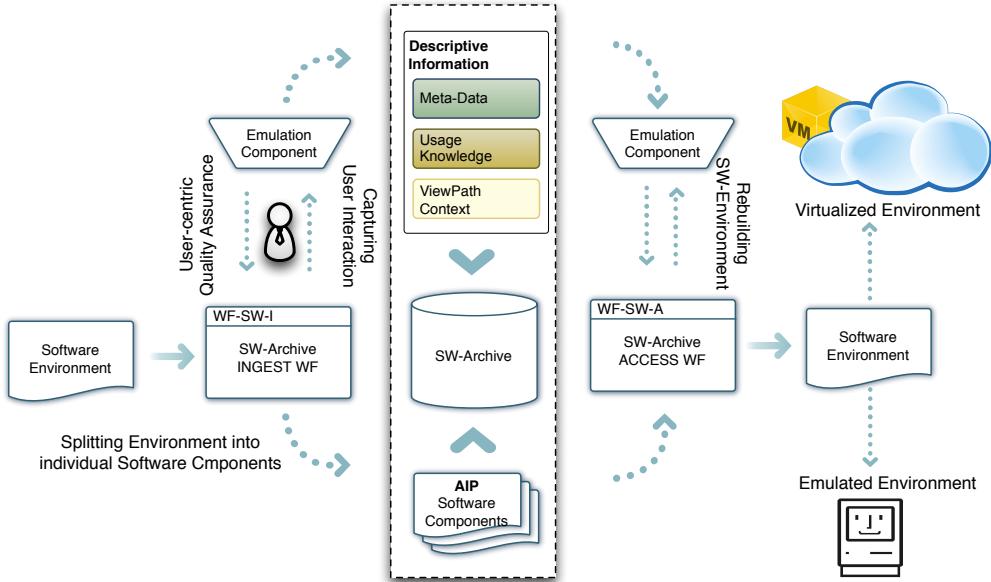


Figure 1: Preservation of complex software environments

on demand. Their approach was successfully tested on different applications in several Windows versions. However, the script language used requires programming skills and in-depth knowledge of the operating system. Within earlier work of the same focus Reichherzer and Brown addressed the creation of emulator images suitable to render Microsoft Office Documents [5].

3. PRESERVATION OF COMPLEX SOFTWARE ENVIRONMENTS

In contrast to a migration strategy, the emulation approach requires a number of additional components and configurations to provide access to digital objects. Thus, independent of which type of emulator is chosen, contextual information of the computer environment is always required. To overcome this gap of missing knowledge, a formalization process is required to compute the actual building blocks for an authentic rendering environment of the digital object.

The VP model describes a system environment starting from the rendering application of the digital object to the description of required software and hardware requirements. If one of these requirements is not met (e.g. hardware components are not available), emulators can be used to bridge the gap between the digital past and future contexts. VPs define an abstract model which can be instantiated by an applicable workflow. To complete the process and compute dependencies technical metadata on various layers is required [3]. Descriptive information needs to be extended, for instance by adding information on required applications, suitable operating systems, and emulators. Additionally, software archiving is required to be able to reproduce complete original environments. Software archives play a vital part in an emulation-centric preservation approach as deprecated software products, legacy hardware drivers, older font-sets, codecs and handbooks for the various programs will become more and more difficult to find.

3.1 Ingest Workflow

Software components need to be preserved and enriched with additional information (meta-data), like operation manuals, license keys, setup how-to's, and usage knowledge. Furthermore, each software component defines its own soft- and hardware dependencies. To ensure long-term access to digital objects through emulation, not only the availability of technical meta-data (e.g. TOTEM entities [1]) are required, but these VPs also need to be tested and evaluated by users aware of the digital object's environment properties and performance. Hence, a defined workflow is required which allows the user to (pre-)view and evaluate the rendering result of each step of VP creation. This can be achieved by providing a framework to perform a structured installation process of a reference workstation. Figure 2 presents a functional flow diagram of the suggested workflow for the addition of new software components to the software archive.

1. The ingest workflow starts with the import of a single software component (WF-I-SW-0). This component might be available through a reference to the digital object already contained in a AIP/DIP container of some digital archive. Otherwise the user is able to upload the object from the local filesystem.
2. In a second step (WF-I-SW.1) the user is able to provide a detailed descriptive information of the object. This description is used as archival meta-data for indexing and search purposes.
3. At workflow step WF-I-SW.2 the user is able to select the software component's hard- and software dependencies. The possible choices are assembled based on already existing knowledge of the software archive or by using external sources. If all required dependencies of the object already exist, the user is able to proceed to workflow step WF-I-SW.3. If the required dependency is not known or not available in the software archive, it must first be ingested into the software archive.

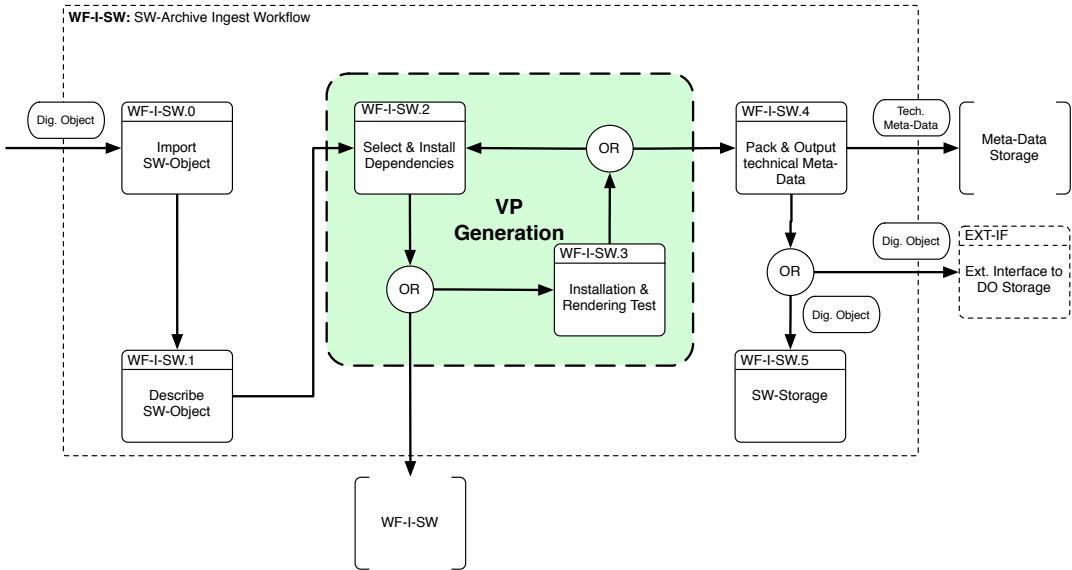


Figure 2: Ingest workflow of a single software package.

archive by using a recursive invocation of the ingest workflow for this missing dependency software component.

4. The options of workflow step WF-I-SW.3 depend on the type of the software component. If it is an operating system, it is run by means of emulation and the user is able to interact with the corresponding environment. If the type of the software component is installable software suitable for a certain operating system (e.g. library, driver, application), it is injected into the system and its installation (either by manual interaction or in an unattended manner) is performed. After the installation is finished, the user is able either to confirm a successful installation or to reject it in case of failure. A successful installation implies automatic extension of the VP for this software component with a new dependency object. Thus after each dependency object is confirmed to have been successfully installed, the VP is extended accordingly until no more dependencies are required for this software component. The resulting VP then represents a suitable manually tested and confirmed rendering environment. If the installation fails due to missing software or hardware dependencies the user has to change the VP accordingly. A repetition of tasks at step WF-I-SW.2 may be required for this.
5. If the user reached step WF-I-SW.4 of the workflow, a suitable VP has been built and a technical meta-data (VP) has been generated. The generated meta-data information might consist not only of the VP but also of user feedback about the quality and/or costs of the produced technical metadata.
6. In a final step the software component is submitted for further processing as SIP to a software archive.

The proposed workflow requires significant manual user interaction and seems costly and time consuming at first

sight. However, regarding preservation of current digital objects, the basic rendering environment is quite stable concerning software and hardware dependencies. Usually the main differences can be found on the top layer of the VP description, i.e. only a few additional steps are required if the software archive already contains suitable VP descriptions of today's common digital objects. The ingest workflow could be further accelerated by employing caching strategies on created software images and by automation of installation tasks.

In order to automate such processes, unattended user interactions with an operating system have led to an interesting possibility of performing automatic dependency installations. So called interactive session recorders are able to record user interactions such as mouse clicks/movements and keystrokes performed by the user during the interaction with an operating system and save them to an interactive workflow description (IWD) file. The interactive session replayers on the other hand are able to read the IWD files and reproduce these actions. Applying this technique to the ingest workflow of the software archive implies recording of all input actions performed by the user during the installation of a software component and saving this information for future purposes. The attractiveness of this approach is that no additional programming must be done in order to automate the installation process, which makes this approach available to a wide range of users and computer systems. Furthermore, the IWD approach is independent of the GUI system used and the underlying operating system [9]. Thus for any successful run of the proposed ingest workflow meta-data (VP) is generated as

$$\begin{aligned} VP_0 &= \langle \text{emulator}, OS \rangle \\ &\dots \\ VP_n &= \langle VP_{n-1}, IWD_n, SW_n \rangle \end{aligned}$$

starting with an emulator / operating system combination which is successively extended by a software component (referenced as TOTEM entity) and the associated installation

and configuration routine.

The combination of base images made for a certain emulator plus a software archive of all required VP software components enriched with knowledge of how to produce a certain original environment (on demand) provides the necessary base layer for the future access of original artifacts. The additional costs in terms of manual interaction during object ingest are able to reduce the long-term preservation planning costs, since only the bottom layer (i.e. emulator) of the VP needs to be taken into account.

3.2 Access Workflow – Rendering of Software Environments

Having a complete VP description for an object is certainly not sufficient for it to be accessed, i.e. rendered. A suitable environment is to be recreated first. In this paper we refer to the process of recreating such an environment by using the term *viewpath instantiation*. A VP is considered as instantiated if the operating system contained in the VP description is started, successfully booted and available for external interaction through the emulated input/output devices. Furthermore, all remaining dependencies defined in the VP for the object need to be installed.

The proposed workflow delegates the task of VP instantiation to a separate abstract service: the emulation component. In order to allow a large, non-technical user-group to interact with emulators an abstract emulation component has been developed to standardize usage and hide individual system complexity. Each Web service endpoint provides a ready-made emulator instance with a remote accessible user interface (currently VNC and HTML5 web output are supported). Furthermore, standard system interaction is available, such as attaching/detaching removable drives (e.g. floppies, CD/DVDs) and attaching hard-drives and images to an emulator. A full description of a distributed emulation setup was presented in earlier work [8].

4. CONCLUSION & OUTLOOK

Emulation becomes a more and more accepted and mature digital preservation strategy to provide access to a wide range of different objects. As it does not demand any modification of the objects over time, the objects do not need to be touched unless requested.

The proposed approach defines a user-centric workflow, which makes use of current user knowledge and thus is able to provide certain guarantees regarding completeness, rendering quality, and non-conflicting dependencies. Furthermore, through a defined framework all interactions between user and computer environment could be observed and recorded. Thereby, not only a more efficient VP instantiation is possible but also knowledge on the usage of certain computer environments and their software components can be preserved. While an emulation approach has technical limitations (e.g. due to external (network) dependencies, DRM, license dongles, etc.), the proposed workflow is able to uncover such issues and indicates risks w.r.t. to long-term preservation.

With the development of a defined work process and associated workflows the groundwork for system integration and automation has been made. With more user experience and feedback, workflow-components suitable for automation could be identified, designed and implemented.

Acknowledgments

The work presented in this publication is a part of the *bwFLA – Functional Long-Term Access*² project sponsored by the federal state of Baden-Württemberg, Germany.

5. REFERENCES

- [1] D. Anderson, J. Delve, and D. Pinchbeck. Towards a workable, emulation-based preservation strategy: rationale and technical metadata. *New review of information networking*, (15):110–131, 2010.
- [2] T. Brody, L. Carr, J. M. Hey, and A. Brown. Pronom-roar: Adding format profiles to a repository registry to inform preservation services. *International Journal of Digital Curation*, 2(2), 2007.
- [3] R. Guenther and R. Wolfe. Integrating metadata standards to support long-term preservation of digital assets: Developing best practices for expressing preservation metadata in a container format. In *Proceedings of the 6th International Conference on Preservation of Digital Objects (iPRES2009)*, pages 83–89, 2009.
- [4] A. N. Jackson. Using automated dependency analysis to generate representation information. In *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES2011)*, pages 89–92, 2011.
- [5] T. Reichherzer and G. Brown. Quantifying software requirements for supporting archived office documents using emulation. In *Digital Libraries, 2006. JCDL '06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on*, pages 86–94, june 2006.
- [6] J. van der Hoeven and D. von Suchodoletz. Emulation: From digital artefact to remotely rendered environments. *International Journal of Digital Curation*, 4(3), 2009.
- [7] R. Verdegem and J. van der Hoeven. Emulation: To be or not to be. In *IS&T Conference on Archiving 2006, Ottawa, Canada, May 23-26*, pages 55–60, 2006.
- [8] D. von Suchodoletz, K. Rechert, and I. Valizada. Remote emulation for migration services in a distributed preservation framework. In *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES2011)*, pages 158–166, 2011.
- [9] D. von Suchodoletz, K. Rechert, R. Welte, M. van den Dobbelaer, B. Roberts, J. van der Hoeven, and J. Schroder. Automation of flexible migration workflows. *International Journal of Digital Curation*, 2(2), 2010.
- [10] K. Woods and G. Brown. Assisted emulation for legacy executables. *International Journal of Digital Curation*, 5(1), 2010.

²bwFLA – Functional Long-Term Access, <http://bw-fla.uni-freiburg.de>.

Practical Floppy Disk Recovery Study

Digital Archeology on BTOS/CTOS Formatted Media

Dirk von Suchodoletz,
Richard Schneider
University Computer Center
Freiburg, Germany

Euan Cochrane
Archives New Zealand
Department of Internal Affairs
Wellington, New Zealand

David Schmidt
RetroFloppy
North Carolina, USA

ABSTRACT

This paper provides a practical example of digital archeology and forensics to recover data from floppy disks originally used by CTOS, now an obsolete computer operating system. The various floppy disks were created during the period between the mid 1980s to the mid 1990s containing different types of text, data and binary files. This paper presents practical steps from two different approaches, the tools and workflows involved which can help archivists and digital preservation practitioners recover data from outdated systems and media. While the floppy disk data recovery was a full success, issues remain in filetype detection and interpretation of non-ASCII data files of unknown or unsupported types.

1. INTRODUCTION

Archives New Zealand and the University of Freiburg co-operated on a data recovery project in 2011 and 2012. The archive received a set of 66 5.25 inch floppy disks from the early 1990s that contained records of a public organization dating back to the mid 1980s. These floppies were not readable using any standard DOS-based personal computer with a 5.25 inch floppy drive attached to it. There was very little information available in the beginning about the contents or technical structure of the floppy disks from the organisation that owned them. Because of the age of the disks and this lack of information about their contents the organisation was eager to retrieve all files that could be read from the disks and get all available information from those files. This is an ideal use case for digital archeology workflows using forensic methods [4, 2, 1] as archives may receive objects quite some time after they have been created (20 years or more later). To be able to recover raw bit streams from obsolete floppies, the archive purchased a special hardware device with the ability to make digital images of the floppy disks. The team from Archives NZ and the University of Freiburg was joined later by a digital archivist from RetroFloppy who had been working on a similar challenge from the same system after

he discovered the discussion about their work on the Open Planets Foundation (OPF) blog.¹

2. STUDY ON DATA RECOVERY

The digital continuity team at Archives NZ thought it would be a great opportunity to demonstrate the practical use of the KryoFlux device, a generic floppy disk controller for a range of original floppy drives offering a USB interface to be connected to a modern computer. In addition, more information on the work required to incorporate it into archival processes was to be gathered and documented.

2.1 First Step – Bit Stream Recovery

The first step in the process after receiving the physical media was to visually examine the disks to find out any technical metadata that was available. The disks had labels that identified them as DS QD 96 tpi disks, which refers to Double Sided, Quad Density, with 96 tracks per inch. A 5.25 inch drive was attached to the KryoFlux which itself was connected to a modern Windows PC using a USB connection. The KryoFlux works by reading the state of the magnetic flux on the disk and writing that signal into a file on the host computer. Different output options are possible: A proprietary KryoFlux stream image formatted file, a RAW formatted file, and an MFM sector (BTOS/CTOS) formatted image file were all created from the disks.

A major component besides the hardware device is the interpretation software to translate the recorded signal into image files that are structured according to various floppy disk-formatting standards. After recovering a few disks it became clear that they were not following any known filesystem standard supported by today's operating systems. Thus it was impossible to directly mount them into the host filesystem and read the files from them. But nevertheless it was possible to analyse the images with a hexadecimal editor. Visual inspection of the resulting data showed that the reading process was producing some meaningful data. Several "words" like *sysImage.sys* were repeated in all readable disk images, thus seeming to represent some structural filesystem data. By searching the internet for this string and others it was possible to deduce that the disks were likely created on a computer running the Burroughs Technologies Operating System (BTOS) or its successor the Convergent Technologies Operating System (CTOS) [5]. Fortunately more in-depth information could still be found on various sites

¹See the discussion on <http://openplanetsfoundation.org/blogs/2012-03-14-update-digital-archaeology-and-forensics>.

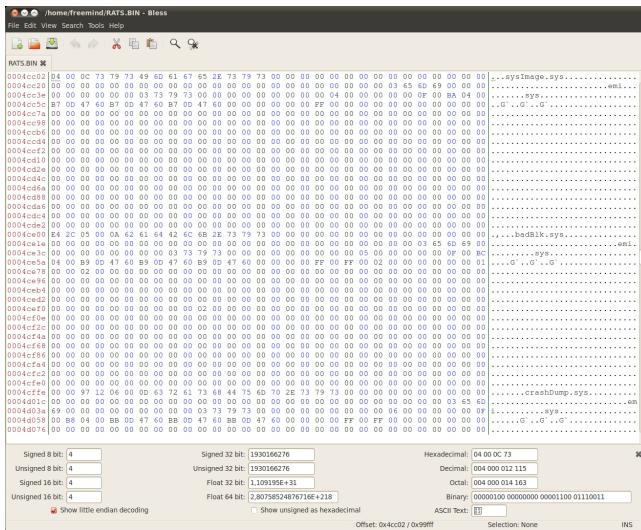


Figure 1: Hexdump image analysis revealing some hints about the original platform

describing the file system. After more research it was concluded that there is currently no software available to properly interpret disks or disk images formatted with this file system aside from the original software and its (obsolete) successors. As there are no emulators available for this system, an emulation approach was not a viable option either. At this point the image investigation was handed over to the computer science department of the Freiburg University to dig into the problem. An application was written to interpret the file system on the disks using the information available on the internet.

2.2 Second Step: Directory Reader

The preservation working group in Freiburg was able to attract a bachelor student for the task to write an interpreter and file extractor for the image files. This is a nice challenge for a computer scientist, as knowledge of operating systems and filesystem concepts are required and could be used practically. There is no demand for a whole filesystem driver, as the image does not need to be mountable on modern operating systems and no new files need to be written. Thus, a bitstream interpreter is sufficient. The Python programming language was used to write a first prototype of the interpreter as there were no performance requirements and it is very well suited for rapid development. By the end of the year a tool was produced that was able to read the filesystem headers and produce directory listings from them (Fig. 2).

In this example the volume header block (VHB) produces a checksum failure, but with the correct File Header Block the simple directory structure is readable. The listing seems to be correct as it reproduces the filenames like sysImage.sys which was readable in the hex editor. With this listing at least some information might be read from the filenames itself. The next stage was the file extraction feature which could extract single files from the image or dump all contained files into a folder on the host system. These could then be inspected further, to gather more knowledge of their original purpose.

```
freemind@blackbox: ~
Datei Bearbeiten Ansicht Terminal Hilfe
freemind@blackbox:~$ ./extract.py RATS.BIN
Info-
Image Name: RATS.BIN
Volume Name: 9.2RATS
Volume Address: 0x0000000000000000
Creation Date: 1985-11-26 12:58:11
Modification Date: 1986-10-01 17:13:59
Directory      Filename      Password      lastModification      Size
SYS            9.2=RATS.RUN      emi        1985-11-26 12:59:20      49664 Bytes
SYS            CAD00] $985:13:54.tmp[DELETED]      1986-10-01 17:13:54      10240 Bytes
SYS            CAD00] $985:13:55.tmp[DELETED]      1986-10-01 17:13:56      10240 Bytes
SYS            MONITORS.PRG      1986-09-15 11:08:18      5825 Bytes
SYS            RAMPACK.DMP      1986-09-15 11:08:18      93535 Bytes
SYS            RAMPACK.mp      1986-09-17 15:22:52      8968 Bytes
SYS            TEST1.pic      MAXINE     1986-10-01 17:08:36      2736 Bytes
sys            bugs.FONT      ul         1986-05-23 14:18:19      8281 Bytes
sys            bugs.RUN      ul         1986-05-23 14:18:19      63489 Bytes
sys            bugzURES      ul         1986-05-23 14:18:27      92 Bytes
sys            joe          1986-09-13 07:19:49      3672 Bytes
sys            joe-old       1986-09-13 07:10:58      3672 Bytes
sys            t1st0909.run      heaven    1986-05-13 23:15      181255 Bytes
sys            trainer       1986-05-13 07:10:59      3672 Bytes
sys            trainer-old    1986-09-13 07:10:59      3672 Bytes
Directories: 2      Files: 15      Total Size: 361932 Bytes
Directories: 2      Summary
Done          freemind@blackbox:~$
```

Figure 2: Python extractor file list output

2.3 Filesystem Interpretation

Of course it was possible to sneak a peek at the probable file contents before, by opening the floppy image file in a hex editor. But this made it very complicated, especially for non-text files to distinguish between file boundaries. Depending on the filesystem used and if fragmentation occurred a single file is not necessarily contained in consecutive blocks on the storage medium. For the preservation and access needs of the Archive and the public institution donating the data, it was not necessary to re-implement the filesystem driver of the old platform for some recent one as most likely nobody will want to write files on floppy disks for this architecture again. But nevertheless a thorough understanding of the past filesystem is required to write a tool that can at least perform some basic filesystem functionality like listing the content of a directory and reading a specific file.

Fortunately the project was started early enough so that all relevant information that was coming from one specific site² on the net was copied locally in time. This site went offline and did not leave relevant traces either in the Internet Archive nor in the publicly accessible cache of search engines. This was a nice example of the challenges digital archaeologists face. Collecting institutions are advised for the future to store all relevant information on a past computer architecture on-site and not to rely on the permanent availability of online resources.

2.4 File Extraction

The extractor program knows the two possible base addresses of the volume header blocks (VHB), as there are an active VHB and one backup VHB defined by the CTOS specification. It selects by the checksum an intact VHB to find the File Header Blocks (FHB). If there is no correct checksum it looks at probable positions for suitable FHB addresses. In the next step the FHB are traversed sequentially to extract the contained file. It will also recover deleted files, files with an inactive header or password secured files. If there are several different, plausible file headers for a file, both files will be saved under different names. After storing the file, its name, directory, password, date and size are displayed as program output. If files cannot be properly

²The site <http://www.ctosfaq.com> went offline permanently, but was replaced later by work done by <http://OoCities.org> archivists at <http://www.oocities.org/siliconvalley/pines/4011>.

identified because of age and wear of the disk images, it will interpret the character encoding as ASCII-encoded strings, which can be extracted easily.

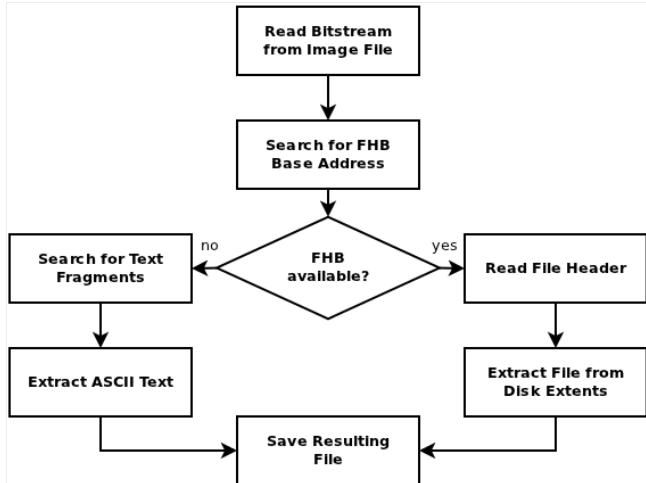


Figure 3: Python extractor for CTOS files and ASCII fragments

3. ALTERNATE APPROACH – FC5025

The RetroFloppy approach used the FC5025 floppy controller, currently available from DeviceSide Data. The FC5025 is a USB-attached circuit board that interfaces to a number of once-common 360 KByte and 1.2 MByte floppy drives. Similar to the KryoFlux device, it reads flux transitions, but exposes much less detail to the user. Designed as a self-contained hardware and software package, it can read multiple disk formats from many different disk systems using a single floppy drive, and includes capabilities to extract entire disk images. Individual files can be extracted from a subset of the supported image formats of the FC5025 driver. In the CTOS case, though, there was no support built in. Fortunately, the FC5025 comes with C-language source code for the formats that are supported. Collaborating with DeviceSide and the other archivists, RetroFloppy wrote code that enabled the FC5025 device to read and interpret the CTOS filesystem, including extraction of individual files (Fig. 4). That support has been contributed back to the vendor for inclusion in future versions of their software package.

As the teams collaborated on filesystem interpretation, differences in strategies emerged and were shared. This ultimately strengthened both approaches. For example, one team felt it was important to extract even deleted files; the other team found significance in file timestamps and passwords. The filesystem interpretation by both teams ultimately relied heavily on available documentation of the CTOS system that would not be discernible by visual inspection. The timestamps, for example, were stored as an offset from May 1, 1952 – presumably a date that was somehow significant to the CTOS project itself, but was not discoverable simply given the disk image data.

4. RESULTS AND EVALUATION

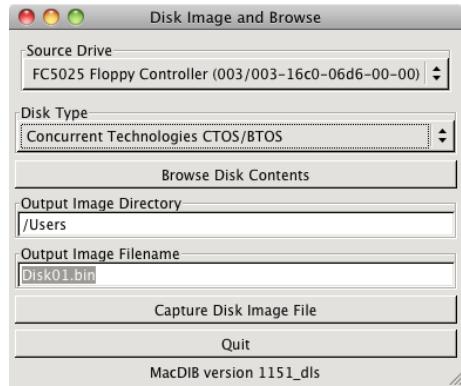


Figure 4: The user interface of the FC5025 solution, showing integrated CTOS support

The first round of the recovery experiment was run on 62 disk images created by the team in New Zealand from the received floppies. In three of those 62 images the File Header Block was unreadable. Two of the failing images had just half the size of the rest of them (320 KBytes instead of 640 KBytes). This issue led to missing file information like file address on the image and file length. For the third failing case it is still unclear why the File Header Block is unreadable. This comes to a total of 59 readable images with a total of 1332 identifiable files in them. The text content of the failing disk images was transferred to a single text file per image. At the moment the issues are being investigated together with the manufacturer of the reading device. It might be possible to tweak the reading process and extract more information to add the missing pieces for the failing images. This might lead to some deeper insight into the procedure and some best practice recommendations.

Filetype	Number of Files
No further recognized data	1635
ASCII text	106
ASCII English text	15
ISO-8859 text	11
XPack DiskImage archive data	7
DBase 3 data file (error)	5
FORTRAN program	2
Lisp/Scheme program text	2
Non-ISO extended-ASCII text	2
8086 relocatable (Microsoft)	1
ASCII C program text	1
Emacs v18 byte-compiled Lisp data	1
MS Windows icon resource (error)	1

Figure 5: Types of the recovered files from the floppy disks by file interpretation.

As the files of the New Zealand test set were mostly of ASCII text, a second set of floppies from the US coast guard was tested at RetroFloppy. This set spanned a longer period and contained many more non-ASCII files. Finally, there were 1889 files successfully extracted from the years 1983 through 1993. Of those, 1789 files with an active file header and 100 deleted files were recovered. The file type detection with the Linux file utility identified most files as "unknown"

binary data" (1635). Eighty-two of them could be identified as CTOS executables by the file extension "run". 838 could be attributed as Word Processor files by extension or manual inspection. Several files got categorized by the *file* utility (5), some of the attributions were simply wrong. In general the results were not widely cross-checked with other utilities as this was not the focus of this study.

5. CONCLUSION

The floppy disk recovery of physically intact media was a full success for both test sets, as it was possible to properly read files from outdated media without any original hardware available. Each disk took approximately five minutes to image and the research and initial forensic work added some additional hours to make up a total of one week of full time work for the initial imaging phase. Less than a month was required for a junior developer to write and test the Python code and less than a week for a seasoned C developer to produce the FC5025 driver code. The future per disk effort should now be very small with the tools available. The most troublesome part of the study was that the only way to understand the file system was to use documentation that has subsequently disappeared from where it was originally found on the internet. This highlights the need for some "body" to independently – not just on some site on the internet – preserve and make available all the system and software documentation for old digital technologies. Without that documentation this kind of work would be much more difficult if not impossible, and at least for the considered platform the documentation is rapidly disappearing.

There are at least two hardware solutions available today, providing an interface between outdated hardware and today's platforms. The KryoFlux device is shipped with proprietary software helping to fine-tune the image extraction. The Device Side USB floppy disk controller is priced very well below \$100 and offers the source code of driver.³ This is definitely a big plus in long-term access. A new controller is currently being developed⁴ that will do similar work to both KryoFlux and DeviceSide FC5025, but is fully open source. So there is clearly interest in the industry in keeping a bridge to older devices open. Both approaches include the ability to extract entire disk images or browse disk directory contents. The two different hardware and software approaches taken here helped to validate and improve the results of both – primarily due to the fact that there were two independent teams working towards the same goal. In the end, the steps taken were the same and would need to be taken by anyone undertaking a project to decode disks of unknown origin:

1. Deduce whatever is possible by visual inspection of the physical media (identifying marks on the disks themselves – bit density, sided-ness, even handwritten clues)
2. Employ a hardware solution to read the bits – KryoFlux is better at configuration "on the fly", DeviceSide FC5025 is simpler to use but requires *a priori* knowledge of and preparation for the format

³The driver page, <http://www.deviceside.com/drivers.html>, gives a list of supported host operating systems and original environments.

⁴The DiscFerret controller, currently under development: <http://discferret.com/wiki/DiscFerret>.

3. Decode the resultant image and retrieve files by following the filesystem conventions of the system that created it.

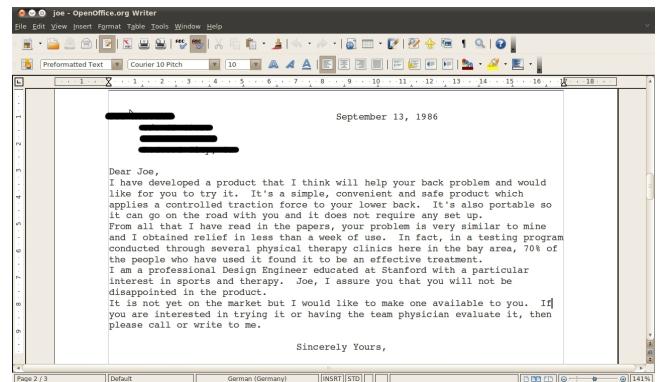


Figure 6: Interpretation of an ASCII text document in OpenOffice

The filetype detection test once again demonstrated the shortcomings of some of the existing tools. It would be great to add some of those files as well as some reference floppy images to a test set of files [3] as especially very old filetypes are under-represented in the existing detection libraries.⁵ The study was brought to a point where some of the files – the ASCII text documents and fragments – could be interpreted, but the content and meaning of the binary data files remains mostly opaque. Another major challenge is the unavailability of software to properly run or render some of the extracted files. Emulation would have been the proper strategy to handle them, but neither a functional emulator nor the required additional software components are available for this computer architecture.

6. REFERENCES

- [1] Florian Buchholz and Eugene Spafford. On the role of file system metadata in digital forensics. *Digital Investigation*, 1(4):298–309, 2004.
- [2] Brian Carrier. *File System Forensic Analysis*. Addison Wesley Professional, 2005.
- [3] Andrew Fetherston and Tim Gollins. Towards the development of a test corpus of digital objects for the evaluation of file format identification tools and signatures. *International Journal of Digital Curation*, 7(1), 2012.
- [4] Matthew G. Kirschenbaum, Richard Ovenden, and Gabriela Redwine. *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Council on Library and Information Resources, Washington, D.C., 2010.
- [5] Edna I. Miller, Jim Croock, and June Loy. *Exploring CTOS*. Prentice Hall, 1991.

⁵The Archive ran a couple of filetype detection experiments on their holdings showing a high failure rate for files dating before the mid 1990ies.

Duplicate Detection for Quality Assurance of Document Image Collections *

Reinhold Huber-Mörk
Intelligent Vision Systems
Safety & Security Department
AIT Austrian Institute of
Technology GmbH
reinhold.huber@ait.ac.at

Alexander Schindler
Department of Software
Technology and Interactive
Systems
Vienna University of
Technology
schindler@ifs.tuwien.ac.at

Sven Schlarb
Department for
Research and Development
Austrian National Library
sven.schlarb@onb.ac.at

ABSTRACT

Digital preservation workflows for image collections involving automatic and semi-automatic image acquisition and processing are prone to reduced quality. We present a method for quality assurance of scanned content based on computer vision. A visual dictionary derived from local image descriptors enables efficient perceptual image fingerprinting in order to compare scanned book pages and detect duplicated pages. A spatial verification step involving descriptor matching provides further robustness of the approach. Results for a digitized book collection of approximately 35.000 pages are presented. Duplicated pages are identified with high reliability and well in accordance with results obtained independently by human visual inspection.

Categories and Subject Descriptors

H.3.7 [Digital Libraries]: System issues; I.5.5 [Pattern Recognition]: Interactive systems

General Terms

Algorithms

Keywords

digital preservation, information retrieval, image processing

1. INTRODUCTION

During the last decade, libraries have been carrying out large-scale digitisation projects, many of them in public-private partnerships with companies like Google or Microsoft, for example, and new digital collections comprising millions of books, newspaper, and journals have been created. Given that each of the single collection items contains up to several hundreds of document images, OCR result files, and other

*This work was supported in part by the EU FP7 Project SCAPE (GA#270137) www.scape-project.eu.

information entities, libraries are facing a paradigm shift in the way how preservation, maintenance, and quality assurance of these collections have to be addressed. Libraries need (semi-)automated solutions that are able to operate on large parts or even on the collections as a whole. Additionally, there are special requirements regarding performance and throughput of the solutions which can be reached by either optimising the time-critical parts of software components or by taking advantage of a distributed software architecture and parallel computing.

In this article, a new approach of document image duplicate detection is presented as a basis for quality assurance in digital library preservation workflows where different versions or derivatives of digital objects have to be maintained and compared to each other. When comparing book pairs, for example, the differences between versions range from variations on the document image level, like additional noise, artefacts, black borders, and more apparent differences due to cropping, page skew, etc., to differences on the object level, like missing or duplicate pages.

Starting with the algorithmic part, there are different aspects of similarity related to document images, including

1. pixel-based similarity, i.e. identity at each pixel, e.g. lossless format conversion, or similarity under lossy compression or radiometrical modifications, e.g. color to greyvalue conversion, color profile adjustment, etc.,
2. similarity under geometrical postprocessing, i.e. scaling, cropping and warping transforms,
3. general similarity induced by independent acquisition under different viewpoint and/or acquisition device and settings.

Figure 1 shows the start of a 730 pages image sequence corresponding to a single book. Starting with the second image a run of eight pages is duplicated from images 10 to 17. Note, that the duplicated images are acquired and post-processed independently. Therefore, the images are geometrically and radiometrically different although showing the same page content.

In general image content comparison is related to visual perception. Perceptual hashing [14, 22], image fingerprinting

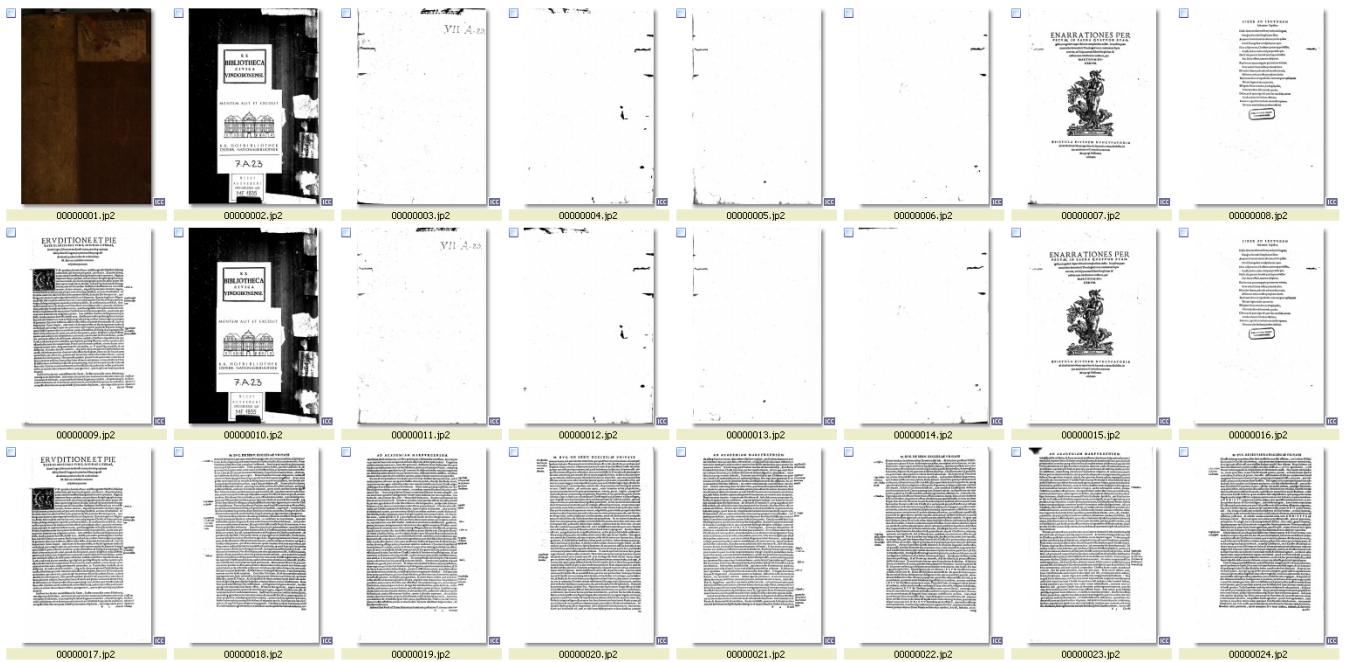


Figure 1: Sample of book scan sequence with a run of eight duplicated pages: images 10 to 17 are duplicates of images 2 to 9 (book identifier is 151694702).

[21] and near-duplicate detection [12, 28] algorithms are related fields. Perceptual similarity, namely structural similarity [25], becomes especially important for comparison of visual content in document images.

Hashing or fingerprinting of images using standard hash functions, like MD5 [18], for example, does only make sense in the narrow domain of bit level image preservation, i.e. if the bitwise representation of the image including all header and formatting information is to be preserved.

The challenges to image processing algorithms can be categorized according to the intensity of preservation actions:

1. The least invasive preservation action for image collections are file format conversions or modifications of the image header information.
2. Preservation actions of moderate intensity are lossy image compression, noise reduction, cropping, scaling and warping transformations, e.g. deskewing.
3. The most invasive modification is completely replacing the representation of an intellectual entity, like the reacquisition of a book in a new scan workflow, for example, possibly involving a different hardware environment and producing differences on the image and/or object level.

Perceptual hashing is interesting especially when significant modifications have been applied to images. Typically, the global characterization of an image, e.g. an individual book page, is obtained by fingerprinting the image with respect to its content. The hashing or fingerprinting function has to be

designed in a way that equal or similar fingerprints are obtained for perceptually similar images, e.g. cropped, denoised or deskewed images, while significantly different fingerprints should be obtained for images with different content, while having similar global characteristics, e.g. color distribution, image dimensions etc.

Global and structural page comparison commonly relies on information or feature extraction from page images. Optical character recognition (OCR) is an established method for information extraction from document images. OCR heavily relies on appropriate page segmentation and adequate font descriptions. Extraordinary page layout, multilingual texts, archaic language and pages containing graphical representations may lead to practical difficulties when taking an OCR based approach. In contrast to web page analysis, where background information regarding the layout can be derived from the document object model (DOM) of the HTML documents, in the case of scanned document images layout information is only achieved by page segmentation. However, good page segmentation results can only be expected if the text regions are clearly structured and the page layout is generally not too complex. Especially for these difficult cases, where reliable background information is not available we suggest a purely image based approach.

Our approach incorporates and extends state-of-the-art computer vision methods for fast object recognition based on the bag of words (BoW) model for condensed representation of image content. We will present a two-stage workflow for image duplicate detection in the context of book preservation which is basically an image fingerprinting approach creating a shortlist of possible duplicates. Spatial verification based on geometrical matching of images followed by structural comparison is then applied to potential duplicates from the

shortlist.

This paper is organized as follows. In Sect. 2 we review related work in document image analysis and computer vision domain. Section 3 presents our approach along with details on the workflow and algorithms. The experimental setup to evaluate our approach and results are presented in Sect. 4. Conclusions are drawn in Sect. 5.

2. RELATED WORK

Image comparison is an applied research area ranging from the inspection of specific objects in machine vision to very general object identification, classification and categorization tasks. Several approaches for the identification of individual objects in large image collections have been proposed in the literature. Typically, approaches in this area make use of local image descriptors to match or index visual information. Near-duplicate detection of keyframes using one-to-one matching of local descriptors was described for video data [28]. A bag of visual keywords [6], derived from local descriptors, was described as an efficient approach to near-duplicate video keyframe retrieval [26]. For detection of near-duplicates in images and sub-images local descriptors were also employed [12].

Image quality assessment can be divided into reference-based (non-blind) [23, 25, 27] and no reference-based (blind) [9, 15] evaluation. It is well known that image difference measures such as taking the mean squared pixel difference does not correspond to the human perception of image difference [24]. To overcome those limitations the structural similarity image (SSIM) non-blind quality assessment was suggested [25]. SSIM basically considers luminance, contrast and structure terms to provide a measure of similarity for overlaid images.

Related work in the field of analysis of document image collections include tasks such as indexing, revision detection, duplicate and near-duplicate detection. Several authors mention that the use of optical character recognition, which is an obvious approach to extract relevant information from text documents, is quite limited with respect to accuracy and flexibility [1, 7, 17].

An approach combining page segmentation and Optical Character Recognition (OCR) for newspaper digitization, indexing and search was described recently [5], where a moderate overall OCR accuracy on the order of magnitude of 80 percent was reported. Page Segmentation is prerequisite for the document image retrieval approach suggested in [1] where document matching is based on the earth mover's distance measured between layout blocks. The PaperDiff system [17] finds text differences between document images by processing small image blocks which typically correspond to words. PaperDiff can deal with reformatting of documents but is restricted as it is not able to deal with documents with mixed content such as pages containing images, blank pages or graphical art. A revision detection approach for printed historical documents [2] where connected components are extracted from document images and Recognition using Adaptive Subdivisions of Transformation (RAST) [3] was applied to overlay images and highlight differences without providing details on the comparison strategy.

The most similar work, compared to our paper is a method for duplicate detection in scanned documents based on shape descriptions for single characters [7]. Similarly to our approach, this approach does not make use of OCR, but, contrarily to our approach, it is based on some sort of page segmentation, i.e. text line extraction.

3. SUGGESTED METHOD

The suggested workflow is shown in Fig. 2, where the digital image collection refers to the set of scanned book images. Note, our analysis is basically applied to individual books independently and what is usually called a document in text image processing refers to an individual page in our setting. The suggested workflow, for which details will be given below, comprises of

1. Detection of salient regions and extraction of most discriminative descriptors using standard SIFT detector and descriptors [13].
2. A visual dictionary following a Bag of Word approach [6] is created from a set of spatially distinctive descriptors.
3. Once the dictionary is set up, fingerprints - visual histograms expressing the term frequency (tf) for each visual work in the corresponding image - are extracted for each image.
4. Comparison of images becomes matching of visual fingerprints and results in a ranked shortlist of possible duplicates.
5. Taking the top-most ranking image gives a fast result for manual post-processing. If one is interested in a more reliable guess the possible duplicate candidates are subject to spatial verification. Spatial verification is realized by descriptor matching, affine homography estimation, overlaying of images and calculation of structural similarity.

3.1 Algorithmic details

In cases of geometric modifications filtering, color or tone modifications the information at the image pixel level might differ significantly, although the image content is well preserved. Therefore, we suggest to use interest point detection and derivation of local feature descriptors, which have proven highly invariant to geometrical and radiometrical distortions [13, 20] and were successfully applied to a variety of problems in computer vision. To detect and describe interest regions in document images we used the SIFT keypoint extraction and description approach. The keypoint locations are identified from a scale space image representation. SIFT selects an orientation by determining the peak of the histogram of local image gradient orientations at each keypoint location. Subpixel image location, scale and orientation are associated with each SIFT descriptor (a 4×4 location grid and 8 gradient orientation bins in each grid cell).

Learning of the visual dictionary is performed using a clustering method applied to all SIFT descriptors of all images, which could become computationally very demanding. As

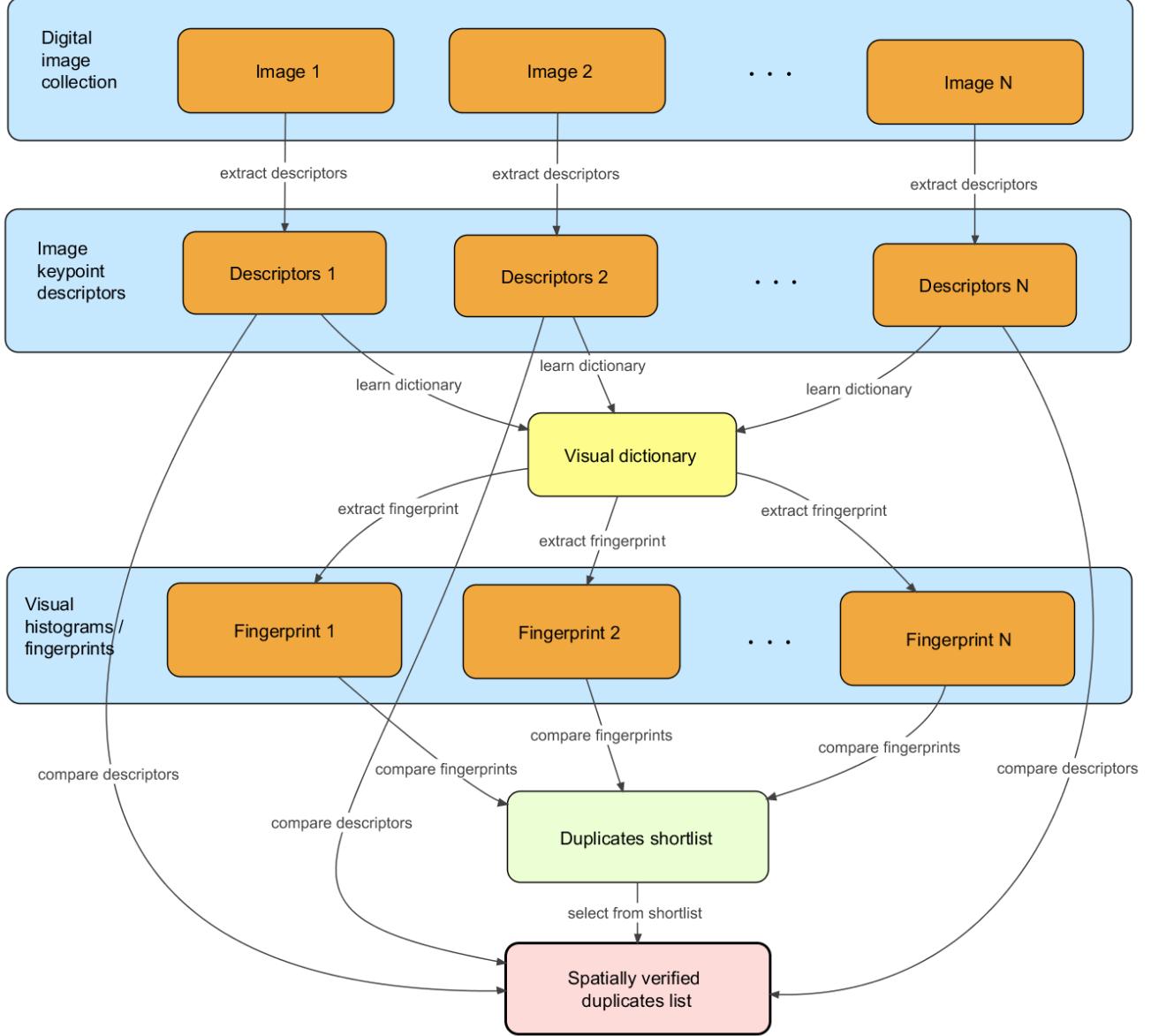


Figure 2: Duplicate detection workflow involving BoW learning, image fingerprinting and spatial verification.

a single scanned book page already contains a large number of descriptors we applied preclustering of descriptors to each image. In contrast to a similar procedure, where all descriptors for all images of the same category are clustered independently and subsequently appended to the BoW [11], we construct a list of clustered descriptors and cluster this list in a second step in order to obtain a dictionary for the whole book. We used k-means for preclustering and final clustering of the BoW. Similar approaches include approximate and hierarchical k-means schemes [16].

Individual terms, or visual keywords, i occur on each page with varying frequency t_i . The visual histogram of term frequencies t_i for an individual book is derived from the BoW representation by counting the indices of the closest descriptors with respect to the BoW. The term frequencies t_i

are represented in its normalized form, i.e. $\sum_{i=1 \dots |V|} t_i = 1$, where V is the set of visual words contained in the visual vocabulary for an individual book.

In order to down-weight the influence of terms occurring in a large number of images and up-weight terms occurring only in some specific images the inverse document frequencies (idf) are optionally combined with the term frequencies [19]. The inverse document frequency idf, in our case better called inverse page frequency, reweights the occurrence of individual visual words on single document image page. We used the common definition of idf for an individual visual word t_i given by

$$t_i^{\text{idf}} = \log \frac{|V| + 1}{(v \in V : t_i \in v) + 1}. \quad (1)$$

The combines tf/idf becomes

$$t_i^{\text{tfidf}} = t_i \cdot t_i^{\text{idf}}. \quad (2)$$

Matching of two visual words t^a and t^b is based on histogram intersection $S_{ab} \in [0, 1]$ given by

$$S_{ab} = \sum_{i=1}^{|V|} \min(t_i^a, t_i^b). \quad (3)$$

At current each page fingerprint is matched against all other page fingerprints. E.g. for a book containing 1000 pages this results in approx. $5 \cdot 10^5$ calculations of vector intersection distances, which could take several minutes on a single core computer.

Spatial verification is based on the established robust matching method called Random Sample Consensus (RANSAC) [8], where corresponding points are randomly drawn from the set of spatially distinctive keypoints and the consensus test is constrained on an affine fundamental matrix describing the transformation between image pairs. The obtained affine transformation parameters are used to overlay corresponding images by warping one image to the other in order to calculate the structural similarity index SSIM.

Spatial verification is computationally very demanding. It was observed that each document image contains 40.000 descriptors on the average. Matching two such images is a bipartite graph matching task requiring $1.6 \cdot 10^9$ computations of the distance between descriptor pairs. On the other hand, spatial matching of images is the most reliable and detailed approach for quality assurance in image preservation. In order to reduce the computational cost on one hand and get access to a more detailed quality assurance method we suggest the following two algorithmic steps:

1. The number of descriptors is reduced in each image by selecting distinctive local keypoints.
2. Descriptor matching is applied to image pairs extracted from the shortlist obtained by image fingerprint matching.

Spatially distinctive local keypoints are obtained by overlaying a regular grid onto each image and selecting the most salient keypoints from local influence regions centered at each grid point. This approach is related to adaptive non-maximal suppression [4], with main the difference that a regular grid and the measure of saliency as used in the Harris corner detector approach [10] is used in our approach. We found that using a grid point number of 2000 delivers sufficiently matching accuracy. Thus, the required number of vector distance computations in spatially matching a pair of images is reduced to $4 \cdot 10^6$. Using a shortlist of moderate size a combined fingerprinting and spatial matching approach becomes feasible.

Combination of fingerprint matching is combined with spatial verification by

$$S_{ab}^{\text{comb}} = S_{ab} \cdot \text{MSSIM}_{ab}, \quad (4)$$

where $\text{MSSIM}_{ab} \in [0, 1]$ is the mean structural similarity index [25].

4. EVALUATION

We evaluated the proposed workflow on a collection of 59 books containing 34.805 high-resolution scans of book pages. Thus, the average number of page images contained in a single book scan was 590. Ground truth data indicating duplicated pages for each book was obtained manually in advance.

The main parameters for the results presented below are summarized as follows. We used standard SIFT features as proposed by [13] providing 128-element vectors. The vocabulary size of the visual BoW was set to 1500 visual words. The number of spatially distinctive keypoints was chosen equal to 2000. The length of the shortlist for spatial verification was 10. All processing was done on greyscale images.

4.1 Comparison of different matching schemes

Using the book with identifier 151694702 and the starting sequence shown in Fig. 1 we compared three query combinations involving

1. visual term frequency histograms only (tf),
2. combined with inverse document frequency (tf/idf),
3. combined with spatial verification (sv).

We calculated the similarity S_{ab} between all image pairs in a book containing N digitized pages. Naturally, there is some level of similarity between text pages due to similar layout, same font etc. Book cover pages have lower similarity to text pages. Finally, duplicated pages show a high similarity. We calculated the maximum similarity found for each image fingerprint when compared to all of the remaining fingerprints

$$S_a^{\max} = \max(S_{ab}), \quad (a, b) \in [1, \dots, N], a \neq b. \quad (5)$$

The considered book shows two runs of duplicates in the scan sequence: page images 2 – 9 are duplicated into page images 10 – 17 and there are nested occurrences of duplicates around page images 108 – 125. We look for local maxim with respect to the scan sequence of Equ. 5 to identify those runs.

Figure 3 shows the S_a^{\max} versus for each image a in the book scan sequence. A sequence of duplicated images starting approximately from image $a = 100$ is visible in Fig. 3 (a) for matching based on tf only. Contrarily, to expected improvement, the tf/idf matching scheme shown in Fig. 3 (b) shows less discrimination for duplicated images. Both methods, tf and tf/idf are not able to identify the duplicated sequence at the start. The reason for this are empty or nearly empty pages, where only a small number of descriptors could be extracted. Finally, Fig. 3 (c) presents tf matching combined with spatial verification applied to a shortlist of length 10.

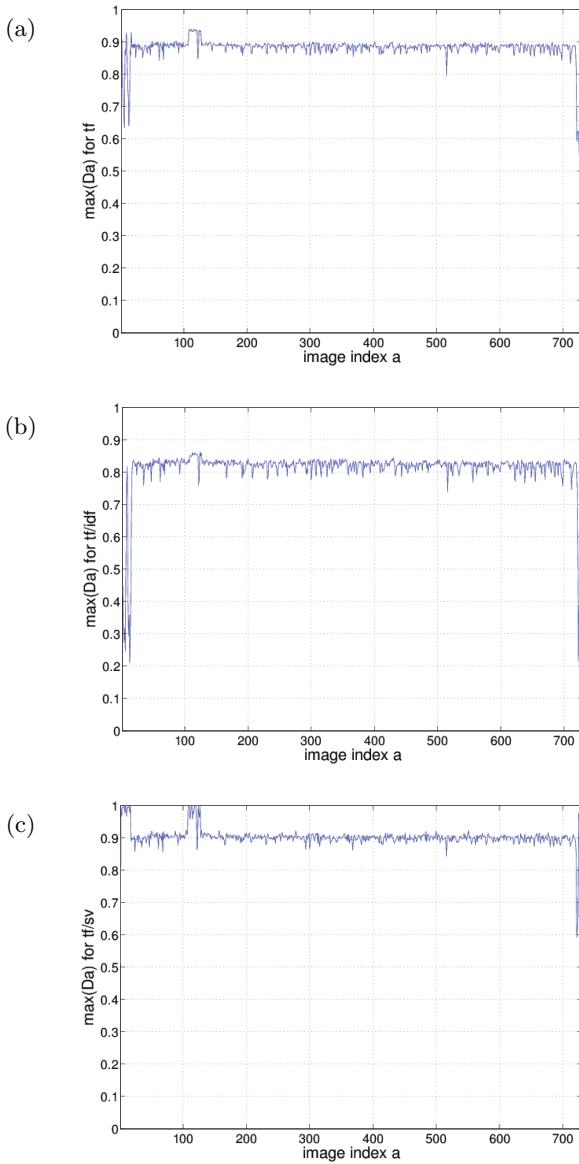


Figure 3: Maximum similarity in duplicate detection using (a) term frequency (tf) only, (b) tf combined with inverse document frequency (idf), (c) tf combined with spatial verification.

Both runs of duplicated sequences of images are visible in this plot.

Remarkably, we observed no advantage using tf/idf compared to the tf matching scheme. The book scan data is characterized by low inter-page variation and the combination with the global idf term seems to lower discriminability for the majority of pages. Therefore, we did not consider the idf term in further experiments. Deeper investigation of this behavior could be topic of future work.

Table 1: Detected duplicates by manual verification and using different image fingerprinting schemes for book 119529601.

Manual detection page	dup.	Automatic detection			
		tf page	tf dup.	tf/sv page	tf/sv dup.
		142	158	142	158
		-157	-173	-157	-173
242	252	242	252	242	252
-251	-261	-251	-261	-251	-261

Table 2: Detected duplicates by manual verification and using different image fingerprinting schemes for book 137274000.

Manual detection page	dup.	Automatic detection			
		tf page	tf dup.	tf/sv page	tf/sv dup.
26	36	26	36	142	158
-35	-45	-35	-45	-157	-173
		264	274	242	252
		-272	-282	-251	-261

4.2 Detailed results for a sample of books

We give a detailed analysis on duplicate detection for a sample of three books. To decide whether an image is a duplicate of another image we applied the following thresholding operation

$$\text{DUP}_a = S_a^{\max} > \left(\text{median}(S_i^{\max}) + n \cdot \text{mad}(S_a^{\max}) \right), \quad (6)$$

where $\text{mad}()$ denotes the median absolute deviation, a robust estimator for the standard deviation

$$\begin{aligned} \text{mad}(S_a^{\max}) &= \text{median}(|S_i^{\max} - \text{median}(S_i^{\max})|), \\ i &= 1 \dots, N, \end{aligned} \quad (7)$$

The parameter $n = 3$ was found experimentally.

We start with analysis of the book with identifier 119528906. Tab 1 shows that both automatic schemes detected two runs of duplicates. The missing first sequence in manual detection was verified to be a real run of duplicate images.

Tab 2 shows the results for the book with identifier 137274000. The tf and the combined scheme detected two runs of duplicates. The ground truth did not contain the second run of duplicates, which was verified to be a real run. In the second sequence there is a gap of a single page image, which caused by the poor quality of the version of the image duplicated at the end of the sequence.

The book with identifier 151694702, also investigated in the last subsection, contains page images occurring three times and even one missing page image. Missing pages could not be detected using our approach. This complicated sequence was identified by both automatic approaches, although it was not found by manual inspection. The tf/sv approach involving spatial verification also detected the duplicate sequence at the beginning of the book. The tf approach was not

Table 3: Detected duplicates by manual verification and using different image fingerprinting schemes for book 151694702.

Manual detection		Automatic detection			
page	dup.	tf page	tf dup.	tf/sv page	tf/sv dup.
2-9	10-17			2-9	10-17
		108	118	108	118
		-111	-121	-111	-121
		112	124	112	124
		-115	-127	-115	-127
		116	124	116	124
		-117	-125	-117	-125
				725	11
				726	3
				727	12
				728	6

able to detect this sequence as is mostly consists of nearly empty pages. Additionally, there were four nearly empty pages at the end of the book which were incorrectly identified as duplicates of the empty pages at the beginning of the book. Table 3 list all sequences of duplicates with their location for different matching approaches.

We will present an heuristics to eliminate the four false detections in the next subsection.

4.3 Results for the whole test corpora

We compare the fast tf matching scheme to ground truth obtained by manual page image inspection. Due to computational complexity, we did not include the tf/sv scheme in this experiment. The decision whether a run of pages is detected by counting the detections DUP_i from Equ. 6 of duplicates locally with respect to the sequence number i . In our case, we used a sequence search range of 10 and threshold on the number of locally detected duplicates of 4. The obtained results are shown in Tab. 4. Interestingly, if there are 2 runs all 2 runs are always detected. In total 53 out of 59 books are correctly treated. There remaining 6 books, which are not correctly classified, are characterized by single runs and atypical image content, e.g. graphical art, high portion of blank pages. The simple thresholding strategy given in Equ. 6 derived from global books statistics seems not appropriate for mixed content.

At current, the ground truth contains only books with runs of duplicates, i.e. there is a detection rate of $53/59 \approx 0.9$. Looking at the number of runs of duplicates, i.e. a total number of duplicate runs of 75 was obtained by manual inspection. Automatic inspection delivered 69 duplicate runs, which results in an accuracy for automatic detection of $69/75 = 0.92$.

Actually, using the automatic method more runs of duplicated images are correctly detected, as already shown in the previous subsection. These additional detection are not shown in Tab. 4.

Further investigation concerning adaptive methods to deal with mixed content and computing strategies to involve spa-

Table 4: Detected runs of duplicates by manual verification and using fast fingerprinting scheme.

Book identifier	Runs		Res	Book identifier	Runs		Res
	M.	A.			M.	A.	
119528906	2	2	ok	119529601	1	1	ok
119565605	1	1	ok	119566804	2	2	ok
119567602	1	1	ok	119572300	2	2	ok
119575003	2	2	ok	119586608	1	1	ok
136403308	2	2	ok	136417009	1	1	ok
136424403	2	2	ok	136432308	2	2	ok
136432400	1	1	ok	136436600	1	1	ok
13646520X	1	1	ok	136465508	1	1	ok
136466203	1	1	ok	136905909	1	1	ok
136975602	1	0	nok	137114501	1	1	ok
137141103	2	2	ok	137141206	1	1	ok
137193905	1	0	nok	137196001	1	0	nok
137203807	1	1	ok	137205804	1	1	ok
137205907	1	1	ok	13721930X	1	1	ok
137220404	1	1	ok	1372237301	1	1	ok
137239607	2	2	ok	137247707	1	1	ok
13727100X	2	2	ok	137274000	1	1	ok
150450702	2	2	ok	150709801	2	2	ok
150711807	1	1	ok	150800701	1	1	ok
150803805	1	0	nok	150816800	1	1	ok
150836306	2	2	ok	150920408	1	1	ok
150930402	2	2	ok	150964102	1	1	ok
150976104	1	1	ok	150976207	1	1	ok
151616508	1	1	ok	151638401	1	1	ok
151671106	1	1	ok	151685609	1	1	ok
151687606	1	0	nok	151694209	1	1	ok
151694702	1	1	ok	151698604	1	1	ok
151699207	1	1	ok	152200609	2	2	ok
152213008	2	2	ok	153936506	1	1	ok
162507508	1	0	nok				

tial verification should further improve the results. Additionally, improved ground truth including the duplicate detection correctly indicated by the automatic method could be derived for future experiments.

4.4 Evaluation in a productive environment

To give an overview on future plans, it is planned to perform an evaluation in a productive environment. First, the accuracy of the book pair comparison is evaluated using an evaluation data set of 50 randomly selected book pairs that will be annotated for that purpose. Second, a large-scale evaluation will be done in order to determine performance and throughput on a distributed system (Hadoop¹). In this context, we compare the runtime of the data preparation and quality assurance workflows on one machine compared to a Hadoop Map/Reduce job running on a cluster with increasing sample size (50, 500, 5000 books) in various steps up to a very large data set (50000 books).

5. CONCLUSION

We have presented an approach for duplicate detection based on perceptual image comparison using image fingerprinting and descriptor matching. The approach reliably indicates positions in the scanned image sequence containing duplicated images for typical text content. We have shown its capabilities on a complicated multilingual and historical book scan data set. Atypical image content, i.e. non-text content,

¹<http://hadoop.apache.org/>

is still an issue to be resolved. Combination with meta-data, such as OCR files and document structure, as well as heuristics incorporating the digitization process, e.g. more detailed information of the scanner operation, through a rule-based system are topics of future research. First heuristics into this direction, i.e. local pooling of duplicate detection events during the scan sequence, were already presented in this work. Further research also includes optimization and deployment of concurrent and parallel computation on the SCAPE platform, especially using the Hadoop Map/Reduce scheme.

6. REFERENCES

- [1] van Beusekom, J., Keysers, D., Shafait, F., Breuel, T.: Distance measures for layout-based document image retrieval. In: Proc. of Conf. on Document Image Analysis for Libraries. pp. 231–242 (April 2006)
- [2] van Beusekom, J., Shafait, F., Breuel, T.: Image-matching for revision detection in printed historical documents. In: Proc. of Symposium of the German Association for Pattern Recognition. LNCS, vol. 4713, pp. 507–516. Springer (Sep 2007)
- [3] Breuel, T.: Fast recognition using adaptive subdivisions of transformation space. In: Proc. of Conf. on Computer Vision and Pattern Recognition. pp. 445–451 (Jun 1992)
- [4] Brown, M., Szeliski, R., Winder, S.: Multi-image matching using multi-scale oriented patches. pp. 510–517. San Diego (June 2005)
- [5] Chaudhury, K., Jain, A., Thirthala, S., Sahasranaman, V., Saxena, S., Mahalingam, S.: Google newspaper search - image processing and analysis pipeline. In: Proc. of Intl. Conf. on Document Analysis and Recognition. pp. 621–625 (July 2009)
- [6] Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV. pp. 1–22 (2004)
- [7] Doermann, D., Li, H., Kia, O.: The detection of duplicates in document image databases. Image and Vision Computing 16(12-13), 907 – 920 (1998)
- [8] Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM 24, 381–395 (June 1981)
- [9] Gabarda, S., Cristóbal, G.: Blind image quality assessment through anisotropy. J. Opt. Soc. Am. A 24(12), B42–B51 (Dec 2007)
- [10] Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. of ALVEY Vision Conf. pp. 147–152 (1988)
- [11] Hazelhoff, L., Creusen, I., van de Wouw, D., de With, P.H.N.: Large-scale classification of traffic signs under real-world conditions. In: Proc. SPIE Electronic Imaging, Conference 8304W Multimedia on Mobile Devices 2012; and Multimedia Content Access: Algorithms and Systems VI (2012)
- [12] Ke, Y., Sukthankar, R., Huston, L.: An efficient parts-based near-duplicate and sub-image retrieval system. In: Proc. of ACM Intl. Conf. on Multimedia. pp. 869–876. ACM, New York, NY, USA (2004)
- [13] Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. of Comput. Vision 60(2), 91–110 (2004)
- [14] Monga, V., Evans, B.L.: Perceptual image hashing via feature points: Performance evaluation and trade-offs. IEEE Transactions on Image Processing 15(11), 3452–3465 (Nov 2006)
- [15] Moorthy, A., Bovik, A.: Blind image quality assessment: From natural scene statistics to perceptual quality. IEEE Transactions on Image Processing 20(12), 3350 –3364 (dec 2011)
- [16] Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proc. of Conf. on Computer Vision and Pattern Recognition (2007)
- [17] Ramachandrula, S., Joshi, G., Noushath, S., Parikh, P., Gupta, V.: PaperDiff: A script independent automatic method for finding the text differences between two document images. In: Proc. of Intl. Workshop on Document Analysis Systems. pp. 585 –590 (Sep 2008)
- [18] Rivest, R.: The MD5 Message-Digest Algorithm. RFC 1321 (Informational) (Apr 1992), <http://www.ietf.org/rfc/rfc1321.txt>, updated by RFC 6151
- [19] Robertson, S.: Understanding inverse document frequency: On theoretical arguments for IDF. Journal of Documentation 60, 503–520 (2004)
- [20] Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. Int. J. of Computer Vision 37(2), 151–172 (2000)
- [21] Seo, J.S., Huitsmu, J., Kulke, T., Yoo, C.D.: Affine transform resilient image fingerprinting. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing. pp. 61–64 (2003)
- [22] Venkatesan, R., Koon, S.M., Jakubowski, M.H., Moulin, P.: Robust image hashing. In: Proc of Intl. Conf. on Image Processing (2000)
- [23] Wang, Z., Bovik, A.: A universal image quality index. Signal Processing Letters, IEEE 9(3), 81 –84 (mar 2002)
- [24] Wang, Z., Bovik, A.: Mean squared error: Love it or leave it? a new look at signal fidelity measures. IEEE Signal Processing Magazine 26(1), 98 –117 (Jan 2009)
- [25] Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600 –612 (April 2004)
- [26] Wu, X., Zhao, W.L., Ngo, C.W.: Near-duplicate keyframe retrieval with visual keywords and semantic context. In: Proc. of ACM Intl. Conf. on Image and Video Retrieval. pp. 162–169. New York, NY, USA (2007)
- [27] Zhang, L., Zhang, L., Mou, X., Zhang, D.: Fsim: A feature similarity index for image quality assessment. IEEE Transactions on Image Processing 20(8), 2378 –2386 (aug 2011)
- [28] Zhao, W.L., Ngo, C.W., Tan, H.K., Wu, X.: Near-duplicate keyframe identification with interest point matching and pattern learning. IEEE Transactions on Multimedia 9(5), 1037 –1048 (Aug 2007)

Audio Quality Assurance: An Application of Cross Correlation*

Bolette Ammitzbøll Jurik
The State and University Library
Victor Albecks Vej 1
DK-8000 Aarhus C, Denmark
bam@statsbiblioteket.dk

Jesper Sindahl Nielsen
MADALGO¹ and
The State and University Library
Victor Albecks Vej 1
DK-8000 Aarhus C, Denmark
jasn@madalgo.au.dk

ABSTRACT

We describe algorithms for automated quality assurance on content of audio files in context of preservation actions and access. The algorithms use cross correlation to compare the sound waves. They are used to do overlap analysis in an access scenario, where preserved radio broadcasts are used in research and annotated. They have been applied in a migration scenario, where radio broadcasts are to be migrated for long term preservation.

1. INTRODUCTION

As part of the SCAPE audio quality assurance work, we have developed a tool called `xcorrSound`, which can be applied in a number of scenarios. The *SCALable Preservation Environments* (SCAPE) project aims to develop scalable services for planning and execution of institutional preservation strategies for large-scale, heterogeneous collections of complex digital objects. To drive the development and evaluation of a number of key outputs from the SCAPE Project, specific real life preservation scenarios have been defined [7].

In this paper we describe two audio preservation cases. The first case is 'access to preserved radio broadcasts for research purposes'. The broadcasts are transcoded for streaming, and an *overlap analysis* is performed to provide a graphical user interface with coherent radio programs.

In the second case the radio broadcasts are to be migrated from MP3 to WAV for long time preservation purposes, and we want to perform *automated Quality Assurance (QA)* on the migrated files. We need to determine if the two audio files (the original and the migrated one) are the same with

respect to their content. This scenario is the SCAPE *LS-DRT6 Migrate mp3 to wav* scenario [5].

There are several ways of designing heuristics that can give some assurance that the migration process went well such as checking if the length is the same before and after the migration. But such 'trivial' measures do not take into account the possibility of just getting white noise as the migrated file, which obviously is a flaw. We will use old and well known techniques from signal processing to catch such errors and report them. The methods we present are easily scalable as well as reasonably reliable.

The algorithms presented in this paper have been implemented in the `xcorrSound` tool package available at [8]. The tool `xcorrSound` finds the overlap between two audio files. `soundMatch` is a tool to find all occurrences of a shorter wav within a larger wav. `migrationQA` is a tool that splits two audio files into equal sized blocks and outputs the correlation for each block (a_i, b_i) , if a and b was the input. The tools all make use of cross correlation, which can be computed through the Fourier transform.

We first present the background for the algorithms in Section 2. Next the algorithms and their applications are described in Section 3. The scenarios are then described in Section 4. In Section 5 we present the experiments for the two scenarios and give the results along with a discussion of these. The non-technical reader should skip Section 2 and Section 3, but for those interested in the implementation details they can be found in those two sections.

2. PRELIMINARIES

The Fourier transform is used in many contexts within digital signal processing. Our algorithms rely on being able to compute the cross correlation of two functions efficiently which can be done using the Fourier transform. Cross Correlation, as the name suggests, gives a measure of how similar two waves are at all offsets (shifting one wave in time and comparing for all time shifts). The peak in the cross correlation is the offset at which the two waves have the highest similarity. This is going to be useful for our algorithms, hence we will recall the mathematical background of these.

DEFINITION 1 (DISCRETE FOURIER TRANSFORM). *Given a sequence of N values x_0, x_1, \dots, x_{N-1} the Discrete Fourier*

¹Center for Massive Data Algorithmics, a Center of the Danish National Research Foundation.

*This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Transform are the complex coefficients

$$X_k = \sum_{n=0}^{N-1} x_n e^{-2i\pi kn/N} \quad (1)$$

for all $k \in \{0, 1, \dots, N-1\}$. We will denote the Fourier Transform of $X = \{x_n\}_{n=0}^{N-1}$ as $\mathcal{F}(X)$.

Straight forward computation of the Fourier transform requires $\mathcal{O}(N^2)$ arithmetic operations, but using the FFT algorithm [11] we can compute it using only $\mathcal{O}(N \log N)$ arithmetic operations.

DEFINITION 2 (DISCRETE CROSS CORRELATION). Let f and g be two discrete complex valued functions, the Cross Correlation is then defined as

$$(f \star g)(t) = \sum_{n=-\infty}^{\infty} \overline{f(n)} \cdot g(n+t) \quad (2)$$

where $\overline{f(n)}$ denotes the complex conjugate of $f(n)$

DEFINITION 3 (DISCRETE CONVOLUTION). Let f and g be two discrete complex valued functions, the convolution is then defined as

$$(f * g)(t) = \sum_{n=-\infty}^{\infty} f(t-n) \cdot g(n) \quad (3)$$

Due to the convolution theorem we can efficiently compute the convolution of two waves if we can efficiently compute the Fourier Transform.

THEOREM 1 (CONVOLUTION THEOREM). Let f and g be two discrete complex valued functions, then we have

$$\mathcal{F}(f * g) = (\mathcal{F}(f) \cdot \mathcal{F}(g)) \quad (4)$$

Proofs of this theorem can be found in any book on Fourier Transforms or signal processing.

We want to compute the Cross Correlation between two wav files. We know that for any real valued function f , $\mathcal{F}(f)(n) = \mathcal{F}(f)(-n)$. Let f and g be the two wav files we want to compute the cross correlation of, and $h(x) = f(-x)$. Note that f and g are wav files thus they can be considered as real valued functions. The Cross Correlation can efficiently be computed:

$$\begin{aligned} (f \star g)(t) &= \sum_{n=-\infty}^{\infty} \overline{f(n-t)} \cdot g(n) = \sum_{n=-\infty}^{\infty} f(n-t) \cdot g(n) \\ &= \sum_{n=-\infty}^{\infty} h(t-n) \cdot g(n) = (h * g)(t) \end{aligned}$$

Now we apply the Convolution Theorem by taking the Fourier Transform and inverse transform on both sides.

$$\begin{aligned} (f \star g) &= \mathcal{F}^{-1}(\mathcal{F}(f \star g)) = \mathcal{F}^{-1}(\mathcal{F}(h * g)) \\ &= \mathcal{F}^{-1}(\mathcal{F}(h)\mathcal{F}(g)) = \mathcal{F}^{-1}(\overline{\mathcal{F}(f)}\mathcal{F}(g)) \end{aligned}$$

One can think of Cross Correlation as taking one of the wave files and sliding it over the other and remember what the best position was so far. Doing it in this way corresponds to computing the Cross Correlation directly from the definition which was $\mathcal{O}(N^2)$ arithmetic operations. Intuitively we are searching for the shift that will minimize the euclidean distance between the two wav files.

3. ALGORITHMS

We have slightly different algorithms for handling the different scenarios but they all rely on efficiently computing the cross correlation of two audio clips. In our implementations of the algorithms we have used the FFTW library [13] for computing the Fast Fourier Transform.

3.1 Computing the Cross Correlation

The input to the Cross Correlation is two periodic functions f and g . When providing a discrete representation of a function as $f(0) = x_0, f(1) = x_1, \dots, f(N-1) = x_{N-1}$, it is assumed that $x_N = x_0$. Because of this, we need to zero-pad the wav files with N zeroes, such that the part that has been shifted “outside” does not contribute anything to the cross correlation value at that particular offset. See Figure 1

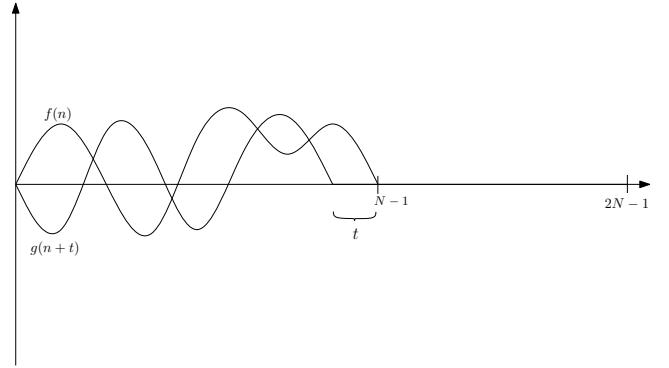


Figure 1: The function g has been shifted t steps in time, and both f and g have been zero padded. Note that from $N-t$ and onwards there is no contribution to the cross correlation, because $g(n) = 0$ for $n \geq N-t$.

If we have two functions f and g given as a sequence of N values indexed by 0 to $N-1$ then we will zero-pad them such that $f(n) = g(n) = 0$ for $n \geq N$. Now we can compute the Cross Correlation as was described in Section 2 because we have a black box (FFTW library [13]) for computing the Fourier Transform.

The Cross Correlation in itself does not provide a measure between $[0, 1]$ describing how much two wav files are alike. We want to normalize it to a value between $[0, 1]$. To do this we divide by $(f \star g)(t)$ by $\frac{1}{2} \sum_{n=0}^{N-t} g(n+t)^2 + f(n)^2$. The resulting value is always less than or equal to 1. The term we divide by can be found efficiently by storing two arrays,

one for each function. The j 'th entry in the array is the sum of the first j values squared. The two prefix sums require only a few arithmetic operations to compute per entry so this will not slow down the computation significantly.

3.2 Overlap algorithm

The input is two wav files where there might be an overlap between the end of the first wav file and the beginning of the second. We are guaranteed that if there is an overlap, it is not longer than a fixed amount of time (2 minutes). We look at the last few minutes of the first wav file and the first few minutes of the second wav file and we compute the cross correlation of these two smaller files. If there is a significant peak, we find it and report that there is an overlap, otherwise we report where the highest peak was, but that it was not very significant. The measure of significance is a value in $[0, 1]$, which is a normalisation of the cross correlation values.

This algorithm was implemented as the `xcorrSound` tool.

3.3 Quality Assurance algorithm

We have two waves and we want to determine whether they are equal or not. Let $X = \{x_n\}_{n=0}^{N-1}$, $Y = \{y_n\}_{n=0}^{N-1}$ be the two audio files. We split these into smaller equal size pieces: $X_0 = \{x_n\}_{n=0}^{B-1}$, $X_1 = \{x_n\}_{n=B}^{2B-1}$, ..., $X_{N/B} = \{x_n\}_{n=N-B}^{N-1}$ (assuming B divides N) and likewise $Y_0 = \{y_n\}_{n=0}^{B-1}$, $Y_1 = \{y_n\}_{n=B}^{2B-1}$, ..., $Y_{N/B} = \{y_n\}_{n=N-B}^{N-1}$. Now we compute the cross correlation for each (X_j, Y_j) pair for $j = 0, \dots, N/B$ and find the peaks. We remember the first peak, and if any of the following blocks' peak position differs by more than 500 samples from the first block's peak we conclude that the files are not similar, otherwise they are similar. We chose B to be 5 seconds worth of samples.

Why is it important to split the files into blocks? The intuition is that if we cross correlate the two files as is, then their similarity may be quite high even if some small parts have very bad correlation which could happen if an error occurred such that there was pure noise for a couple of seconds somewhere in the wav file.

3.4 Analysis

The quality assurance algorithm runs in $\mathcal{O}(N \log B)$ time since we split the N samples into N/B blocks of size B then each cross correlation will take $\mathcal{O}(B \log B)$ time to compute, hence the execution time follows. We, however, care a great deal about the constants. For every block we need to perform three Fourier transforms and roughly $4B$ multiplications and divisions. Notice that the unit for N and B is samples. One way to speed up the tools is to simply have lower sample rates and then there will be a trade-off between the quality of the results and the sample rate. The intuition is that radio broadcasts probably do not need 48kHz sample rate and if we have two wave files that are very similar then down sampling should not change the similarity significantly.

We are also interested in the robustness of the migration algorithm. The primary question is, how degraded is the material allowed to become when migrating? Cross Correlation is quite robust wrt. artifacts (eg. extra noise in the background) appearing in addition to what was supposed to be in the result file. By robust, we mean that the algo-

rithm will likely still find the correct offset, but the value of the match decreases as more noise is present. One way to solve degradations like this is either to output some of the blocks that had a low match value for later manual checking or do experiments on degraded signals and fix a parameter that can decide whether the migrated file has an acceptable quality. The last method has the disadvantage that when migrating the same file through a number of intermediate steps it will (maybe) be unrecognizable in the end, though every intermediate step was within the acceptable parameters. Think of this as making a copy of a copy of a ... of a copy of a newspaper article.

4 SCENARIOS

Both the access scenario and the migration scenario are well known in relation to digital preservation [19]. Transcoding or migrating audio and video for access is done as the 'preservation master' is usually too big a file to share with users, maybe it cannot be streamed online, and the "popular" online access formats change [12]. The overlap analysis is relevant in our context as audio broadcasts were recorded in two hour chunks with a few minutes of overlap, and we want to find the exact overlap to make an interface to the broadcasts without strange repetitions every two hours. Migration of audio from MP3 to WAV is done primarily as the WAV is the IASA (International Association of Sound and Audiovisual Archives) recommended preservation format [10].

4.1 Finding Overlap

In connection with the LARM project WP2, the *overlap analysis* issue arose. The LARM project [4] is a collaboration between a number of research and cultural institutions in Denmark. The project provides research data and meta data to a digital infrastructure facilitating researchers' access to the Danish radio-phonic cultural heritage.

The addressed and problematic collection is Danish radio broadcast from 1989 till 2005 from four different radio channels. The recordings were made in two hour chunks on Digital Audio Tapes (DAT), and were recently digitized. Our library got MP3 (and not WAV) copies of these files primarily due to storage limitations. High resolution WAV-files also exist within the broadcasting company. The MP3 files have sampling rate 48 kHz and bit depth 16. The collection is roughly 20 Tbytes, 180000 files and 360000 hours.

In order not to lose content originally, one tape was put in one recorder and a few minutes before it reached the end, another recorder was started with another tape. The two tapes thus have a short overlap of unknown duration, as do the digitized files.

The task is to find the precise overlaps, such that the files can be cut and put together into 24 hour blocks or other relevant chunks correctly.

4.2 Migration QA

The Danish radio broadcast MP3 files are also addressed in the SCAPE *LSDRT6 Migrate mp3 to wav* scenario [5]. They are part of the Danish cultural heritage the Danish State and University Library preserves. They are used as examples of a very large MP3-collection well knowing that original WAV

files actually exist for this collection. We have other collections in both MP3 and other compressed and/or older audio formats that could and should be migrated to WAV at some point in time but chose to work with the same collection for the two scenarios to ease the work. This means that the library would like to migrate the files to WAV (BWF) master files, as is the IASA recommendation [10]. This format has been chosen as the preferred preservation format as this is a raw format, which needs less interpretation to be understood by humans, and is also a robust format. The actual migration is done using FFmpeg [1]. The decompression presents a preservation risk in itself, which is why keeping the original MP3s and performing quality assurance (QA) on the migrated files is recommended.

The QA is done in a number of steps. The first step is validation that the migrated file is a correct file in the target format. We currently use JHOVE2 [3] for this validation.

The second step is extraction of simple properties of the original and the migrated files, and comparing these properties to see if they are 'close enough'. We currently use FFprobe to extract properties. FFprobe is a multimedia streams analyzer integrated in FFmpeg. The properties that are checked are sampling rate, number of channels, bit depth and bit rate.

We could add a third step of extracting more advanced properties using a tool such as e.g. Cube-Tec Quadriga Audiofile-Inspector [15] and comparing these properties. Note however that tools such as Cube-Tec Quadriga Audiofile-Inspector do not compare content of audio files, but rather provides an analysis of a single audio file. We are evaluating significant properties, property formats, property extractors and comparators for possible addition to the workflow.

We have run the migration, validation and property comparison workflow on some of the Danish radio broadcast MP3 files creating a small test set for further QA. The workflow is written as a Taverna [9] workflow and is available on myExperiment [16]. The workflow used SCAPE web services which are set up locally. The used SCAPE web services are the FFmpeg, JHOVE2 and FFprobe web services defined in the scape GitHub repository [6]. Comparison of the extracted properties is done with a Taverna bean shell. The workflow input value is a fileURL containing a list of input MP3 URLs. The output is a list of Wav fileURLs, a list of validation outputs (valid / not valid) and a list of comparison outputs (properties alike / not alike).

The test set contains 70 Danish radio broadcast MP3 files. The workflow was run on a test machine with an Intel(R) Xeon(R) CPU X5660 @ 2.80GHz processor and 8GB RAM running Linux 2.6.18 (CentOS). The workflow finished in approximately 5 hours and 45 minutes. This means we have a performance of almost 5 minutes pr file. Earlier tests have shown that the most expensive components in the workflow is the FFmpeg migration and the JHOVE2 validation, while FFprobe characterisation and property comparison is relatively cheap [18].

We note that the Danish Radio broadcasts mp3 collection is 20 TB and around 180000 files. This means that running

the basic workflow migrations sequentially on the test machine would take more than 600 days. We do however plan to improve that significantly by using the Scape execution platform instead of doing the migrations sequentially on just one server.

Another related scenario is that The Danish State and University Library have a very small collection of Real Audio (200 files) that are planned to be migrated to wav. The actual FFmpeg migration needs adjustment and we need to find another independent implementation of a Real Audio decoder, but the rest of the workflow as well as the algorithms presented in this paper can be applied to this issue directly.

5. EXPERIMENTS

5.1 Overlap Analysis Tool Use

We have already used the overlap tool on real data sets. The **xcorrSound** tool is used to find the overlaps. The solution must consider

- Some recordings (files) may be missing.
- Noise at both ends of the sound files. Can be both silence and changing to a different station.
- The sound recording may have been started up to 23 minutes early.
- There must be a quality assurance to show that the transformation was successful. The tool used for this QA is also the **xcorrSound** tool. The success criteria are:
 - Good overlap measurement. QA check match value of at least 0.2
 - Length of resulting file is **not** checked, as above check also catches these cases.

The overlap match is done by a script, which first cuts a short interval (1 second) of either end of the files, as this is often noise related to the start or finish of recording, see Fig. 2. Then a time interval of 10 seconds in the second file is cut for later QA analysis. The **xcorrSound** tool is now run on 6 minutes of the end of the first file and the beginning of the second file. The output is a best match position and best match value. Using the best match position, the **xcorrSound** tool is run a second time on the 10 second time interval cut for QA. If the best match value is acceptable, the files are cut and concatenated at the best match position.

The results were initially all checked manually to estimate the acceptable values for the best match. The results where the best match value is not acceptable, are checked manually and the script is tweaked to hopefully provide matches.

5.1.1 Results

Our data set consisted of one month of radio broadcasts recorded in 2 hour chunks. The goal was to cut them into 24 hour chunks instead. The **xcorrSound** tool worked very well. We found that when doing the QA check, if the value produced was below 0.1 there was an error and if the value

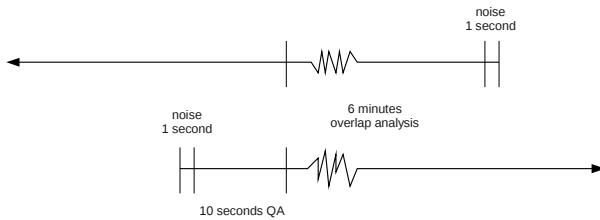


Figure 2: Overlap Analysis

was above 0.2 it was safe to assume the process went correct. We found several actual errors in the content using this tool. Examples include that one file simply contained a wrong radio show (may have happened if a channel was changed), several files in a row were identical, hence they would not overlap in the end and an error would be produced or there was a lot more overlap than the promised few minutes - up to 25 actually. All these errors in the data set was caught and reported. The tool of course only tells when two files do not overlap and the actual reasons have to be manually found. QA values that lie in the range 0.1 - 0.2 are the ones that we are not quite sure of and we would do manual quality assurance on these. However It is rare that the QA values lie in that range and most commonly the QA match is above 0.7.

5.1.2 Discussion

The `xcorrSound` tool has been used and the results were quite good. We found several errors in the collection that we can now correct. As can be seen we have a nice structure on the values of the QA match. We have found that by doing experiments and trying to listen to the broadcasts and comparing with the QA match values we can now run through a large collection and do automatic quality assurance because we have determined the different intervals we can trust for the QA values. We know that when a QA value is below 0.1 there is almost surely an error and when the QA value is above 0.2 there is not an error.

5.2 Migration QA Tests

In order to test the `migrationQA` tool we needed a data set. The test data set contains 70 two-hour radio broadcast files which were migrated using FFmpeg[1], see Section 4.2. The average file size of the original mp3 files is only 118Mb, but the migrated wav files are approximately 1.4Gb. Three of them were replaced by a randomly generated file with a 'correct' wav-header, such that the `migrationQA` tool was able to process them. We assume that checks such as correct header information are performed before invoking the `migrationQA` tool. Five of the remaining 67 files were kept intact except for a few seconds a few places within the file which were replaced by randomly generated bytes. The other 62 files were kept as they were after migrating through FFmpeg. We have an inherent problem using this data set because it is quite artificial. We imagine that the data set contains errors that *might* occur during a migration, but we have no basis for this as we have never seen any erroneous migrations. To use the `migrationQA` tool we need to 'play' or interpret the files, just as a human needs to 'play' or interpret an mp3 file to

hear the sound. We currently use MPG321[2] to 'play' the original mp3 files. MPG321 is an independent implementation of an mp3-decoder, thus independent from FFmpeg, which was used to migrate the files. The migrated files are already in wav format and are used directly.

The *migrationQA SCAPE Web Service including MPG321 decoding Workflow* [17] on myExperiment takes an mp3 file and a wav file as input. The mp3 file is then played into a temporary wav file using MPG321, and the temporary file and the input wav file are compared using the `migrationQA` tool. This workflow however only works on one pair of files.

We have tested the tool on a list of 70 pairs of mp3 and migrated wav test files using a bash script. The *migrationQA including MPG321 workflow bash script* was run on a test machine with an Intel(R) Xeon(R) CPU X5660 @ 2.80GHz processor and 8GB RAM.

5.2.1 Results

The script ran for 4 hours and 45 minutes. This gives us a performance of just over 4 minutes pr. file. This is roughly equally divided between the MPG321 migration and the `migrationQA` comparison.

In total there were 12 reported errors, which is 4 more than we expected. All the files that were supposed to be found during this QA check were found, so we only have some false positives left (or false negatives depending on your view). We investigated the additionally reported errors. The 'limit' of 500 samples difference from the first block may in fact be too low. On one pair of files the best offset was 1152 samples during the first 6850 seconds of the file (00:00:00-01:54:10) but during the remaining part of the file it changed to having the best offset at 3456 samples and a cross correlation match value of nearly 1 (0.999-1.0).

5.2.2 Discussion

The fact that partly through the file the best offset changed suggests that either one of the converters has a bug or there were some artifacts in the original mp3 file that is not following the standard and thus they simply do not recover from this in the same manner. Of course when there are 48000 samples/second we cannot hear the difference between an offset on 3456 and 1152 (4.8 milliseconds). Now the question is as much political as it is implementational. Was the migration correct or was it not? Obviously one can argue that since we cannot hear any difference between the two files, the migration went as it should. On the other hand, one of the files we ran the `migrationQA` program on must have had some errors, if we accept that one of the files must be a correct migration. Ultimately the question is up to the definition of a correct migration, which is a subject we have carefully avoided in this paper. One solution is to let the `migrationQA` program take a parameter that decides how much difference from the first block is allowed, rather than fixing a magic constant of 500 samples. Another solution is to try to identify what exactly is happening inside the migration tools (FFmpeg and MPG321) to find out why they differ and check if one of them has a bug or if it was in fact the original mp3 file that did not follow the standard.

One might argue that the `migrationQA` program is as much

a validation tool of other programs that migrate audio files to wav to check if they agree as it is a Quality Assurance tool for migrated files. This happens when we accept one migration tool to be correct and then try migrating a lot of files using that tool and another we want to test correctness of. If they agree on the output, then we can have some confidence the other migration tool is correct as well.

In this paper we had two ways of migrating an mp3 file to wav, but we were unsure whether any of them were correct. If we assume that the migration tools are independent implementations this should intuitively provide some assurance that they do not have the same error (if they have any). Hence, if they agree on the output we have some confidence that the migration went as it should. The question is, if it is a reasonable assumption that they do not have the same error if they are independent implementations. They are after all implementing the same algorithm, which likely has some parts that are non trivial to implement and others that are trivial.

We care a great deal about efficiency, and just over 4 minutes per file is at the moment acceptable. The algorithm is not easy to make parallel but it is easy to have several instances of the same algorithm running. This is a feasible solution because the algorithm can be implemented to use a limited amount of memory. All that needs to be in memory at any point is the match value and offset of the first block, the current block being processed and some buffers for speeding up the I/O. Our implementation uses roughly 50mb memory when running. If we have a machine that can run multiple instances of the program, it might be the I/O operations that become the bottle neck of the program.

6. CONCLUSION AND FURTHER WORK

We presented algorithms for doing Quality Assurance on audio when migrating from one file format to another. We also gave an algorithm to eliminate overlap between audio files such that potential listeners do not need to hear the same bit twice. The experiment for QA showed that the tool works well on the constructed input. Since we do not have any data where the migration goes bad we cannot speak to how good the tool actually is, but we believe that it will work very well. The experiment also showed that there is not one single algorithm that will fit all. It might be necessary to fiddle with parameters depending on the data set being processed. Further work in this area is to try to develop even faster algorithms and develop better metrics for comparing audio. We used the Cross Correlation metric, which is a relatively expensive metric to compute, perhaps there are cheaper ones that work just as well or more expensive ones that can give better guarantees. For doing the overlap analysis we could possibly have adopted a finger printing scheme (such as [14]) that would have worked just as well, though that solution is a lot more complex than our suggested approach. The technique of applying cross correlation is general and might have application elsewhere for doing QA. It is worth investigating if we can reuse the same ideas for other areas as well.

Acknowledgements

Thanks to Bjarne Andersen, Henning Böttger and Asger Askov Blekinge for all their work on the experiments and help with the paper.

7. REFERENCES

- [1] FFmpeg (2012), ffmpeg.org
- [2] Homepage of mpg321 (2012), mpg321.sourceforge.net
- [3] JHOVE2 (2012), jhove2.org
- [4] LARM audio research archive (2012), www.larm-archive.org/about-larm/
- [5] LSDRT6 migrate mp3 to wav (2012), wiki.opf-labs.org/display/SP/LSDRT6+Migrate+mp3+to+wav
- [6] SCAPE project repository (2012), <https://github.com/openplanets/scape>
- [7] SCAPE scenarios - datasets, issues and solutions (2012), wiki.opf-labs.org/display/SP/SCAPE+Scenarios++Datasets%2C+Issues+and+Solutions
- [8] Scape xcorr sound tools (2012), <https://github.com/openplanets/scape-xcorr sound>
- [9] Taverna workflow management system (2012), taverna.org.uk
- [10] Committee, I.T.: Guidelines on the production and preservation of digital audio objects. standards, recommended practices and strategies, iasa-tc 04, www.iasa-web.org/tc04/audio-preservation
- [11] Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex fourier series. Mathematics of Computation 19(90), 297–301 (1965)
- [12] Elsheimer, S.: Introduction to transcoding: Tools and processes (2011), www.prestocentre.org/library/resources/introduction-transcoding-tools-and-processes, presentation at Screening the Future 2011
- [13] Frigo, M., Johnson, S.G.: FFTW library (April 2012), <http://www.fftw.org/>
- [14] Haitsma, J., Kalker, T.: A highly robust audio fingerprinting system. In: Proceeding of the International Symposium on Music Information Retrieval (ISMIR) (2002)
- [15] Houpert, J., Lorenz, T., Wieschollek, M.: Quadriga - audiofile-inspector - cube-tec international (August 2012), <http://www.cube-tec.com/quadriga/modules/audiofileinspector.html>
- [16] Jurik, B.: Workflow entry: Migrate mp3 to wav validate compare list to list (May 2012), www.myexperiment.org/workflows/2914.html
- [17] Jurik, B.: Workflow entry: migrationqa scape web service including mpg321 decoding workflow (March 2012), <http://www.myexperiment.org/workflows/2806.html>
- [18] Pitzalis, D.: Quality assurance workflow, release 1 & release report (March 2012), <http://www.scape-project.eu/deliverable/d11-1-quality-assurance-workflow-release-1-release-report-draft>
- [19] Wright, R.: Preserving moving pictures and sound. dpc technology watch report 12-01 march 2012. Tech. rep., The Digital Preservation Coalition (DPC) (2012)

Evaluating an Emulation Environment: Automation and Significant Key Characteristics

Mark Guttenbrunner
Secure Business Austria
Vienna, Austria
mguttenbrunner@sba-research.org

Andreas Rauber
Vienna University of Technology
Vienna, Austria
rauber@ifs.tuwien.ac.at

ABSTRACT

Evaluating digital preservation actions performed on digital objects is essential, both during the planning as well as quality assurance and re-use phases to determine their authenticity. While migration results are usually validated by comparing object properties from before and after the migration, the task is more complex: as any digital object becomes an information object only in a rendering environment, the evaluation has to happen at a rendering level for validating its faithfulness. This is basically identical to the situation of evaluating the performance in an emulation setting.

In this paper we show how previous conceptual work is applied to an existing emulator, allowing us to feed automated input to the emulation environment as well as extract properties about the rendering process. We identify various significant key characteristics that help us evaluate deviations in the emulator's internal timing compared to the original system and how we can find out if the emulation environment works deterministically, an important characteristic that is necessary for successful comparison of renderings. We show the results of rendering different digital objects in the emulator and interpret them for the rendering process, showing weaknesses in the evaluated emulator and provide possible corrections as well as generalized recommendations for developing emulators for digital preservation.

1. INTRODUCTION

Preserving digital information for the long term means to adapt it to be accessible in a changed socio-technological environment. But applying a preservation action like migration or emulation on a digital object changes elements in the so-called view-path. This includes not only the object but also secondary digital objects needed to render it, i.e. the viewing application, operating system, hardware or rendering devices. To strengthen the trust in these digital preservation actions we have to validate the rendered form of the object (where "rendering" means any form of deploying an information object, being rendering it on a screen or on

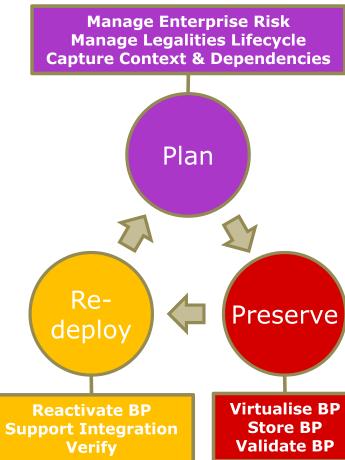


Figure 1: Process for Digital Preservation of Business Processes (BP) in TIMBUS.

any other form of output, including acoustic, physical actuators, output on data carriers or TCP/IP ports, etc.) Thus, migration and emulation, usually perceived to be drastically different approaches in digital preservation, actually become rather similar in their principles of evaluating the rendering of the object.

The principles are the same: devise a way to capture information from a rendering environment (which we will, without limiting its general applicability, refer to as "emulator" for the remainder of this paper, and where we will use a concrete emulator as a compact form of a system comprising key elements of the rendering environment providing access to a digital object). We devised a formal framework to evaluate the rendering of digital objects in [8] that is applicable to all kinds of objects, from static files to dynamic processes. In this paper we will validate this framework and show a detailed evaluation of the rendering process.

Evaluating digital preservation actions performed on digital objects becomes a necessity when doing preservation planning to support the decision for the most suitable strategy and tool to perform the action. Similarly the validity of the preservation action has to be checked when preserving the object by executing the preservation action on it, as well as when validating the object once its re-deployed for future

execution in a new environment. The different stages as defined in the TIMBUS¹ project are shown in Figure 1, and are explained in detail in [1]. To compare the renderings in these different stages of an object’s life cycle, we have to extract characteristics about the rendering process as well as data rendered during this process from the environment. But to reliably compare two different renderings of a digital object it is necessary to avoid side-effects from manual input and other non-deterministic aspects, so we need to automate the evaluation process.

In-depth information about the rendering process is only known inside of the rendering environment. In the case of emulation this is inside the emulator. Based on this we argue that emulators used for digital preservation have to offer functionality to support their evaluation. Based on the theoretical work on the features we would expect emulators to offer [8], we show how we implemented some of these in an existing emulator. We also show how these features are used to automate input to the emulation environment to support automated and repeatable testing uncoupled from exactly timed manual user input. We describe significant key characteristics that we extract from the log files created by the emulator about the rendering processes of various different digital objects. These characteristics are analyzed and used to improve the emulator.

While applicable to all kind of preservation actions, in this paper we focus on emulation. We picked an open-source emulator of a home-computer environment as an example of sufficient but still manageable complexity. Using two types of applications with different characteristics and complexity, namely a game as well as a simple, early business application allowing the management of income and expenses, we will validate the key characteristics and feasibility of the proposed approach, and show how these extend to more generic business or eScience processes of generic object renderings.

This paper is structured as follows. First we provide related work on the evaluation of digital preservation actions. Then we give a brief overview of the emulator we chose for evaluation in Section 3. For the remainder of the paper we present the theoretical work on evaluation and how it is implemented in the emulator: We first show in Section 4 how we implemented an event-log. Then we show in Section 5 how we used this log for automated execution of the emulator. In Section 6 we describe how the created logs can be used to extract characteristics about the rendering process and how those can be used for evaluating an emulator. In Section 7 we describe the experiments we performed on different digital objects in the emulator and describe the findings in the rendering logs. Finally, we show our conclusions and give an outlook to future work.

2. RELATED WORK

Choosing the right preservation action for digital objects is a challenging task. To give the team responsible for performing digital preservation activities a certain level of certainty about the digital preservation actions performed, it is necessary to validate the effects of these actions on the significant properties of digital objects.

¹<http://timbusproject.net/>

In [2] a preservation planning workflow that allows for repeatable evaluation of preservation alternatives, including migration and emulation strategies, is described. This workflow is implemented in the preservation planning tool *Plato* [3]. As part of the preservation planning automatic characterization of migrated objects can be performed. Tools like Droid [5] are used to identify files. Migration results can be validated automatically supported by the eXtensible Characterisation Languages (XCL) [4]. The original and migrated objects are hierarchically decomposed and represented in XML. These representations can be compared to measure some of the effects of the migration on the digital object. It is, however, not immediately obvious if all the significant properties of a digital object are sufficiently reproduced once it is rendered in a new rendering environment. This new rendering environment can be either different software used to render the migrated file or, in the case of emulation, a new environment in which the original file is rendered.

Comparing rendering results to evaluate the outcome of a rendering process was proposed in [12] as separating the information contained within a file from the rendering of that information. The information stored in the file can, for example, be the coordinates of text or descriptive information about the font to use while the rendering displays the text on a specific point on the screen and uses either a font built into the system or a font stored within the file, which in turn is also rendered in a way specific to the application used for displaying the document. This is described as the *look & feel* aspect of an object. In [9] case studies of interactive objects comparing the rendering outcomes of different rendering environments using the aforementioned characterization language XCL on the level of screenshots of renderings are presented.

Most approaches to evaluate the validity of emulators as a preservation strategy are currently based on manually reviewing the emulation results. In the CAMiLEON project [10] users compared objects preserved by different strategies including emulation. The emulated environments were evaluated by the users as subjective experience with the preserved digital object. A case study to compare different approaches to preserve video games with one of the approaches being emulation was also reported in [6] on a human-observable and thus also to some extent subjective level.

A manual comparison of original and emulated environment is a very time consuming process, that would have to be repeated whenever a new emulator or a new version of an emulator is introduced in an archive due to the necessity of a digital preservation action, e.g. if the hardware platform used for the previous emulator gets obsolete or if any other element in the viewpath (including any system changes on the host environment running the emulator) or on the level of output devices used for the rendering of an object that may have an effect on the result of performing/rendering an information object, change.

In [8] we presented a framework which allows one to determine the effects of an emulated environment on the rendering of objects in a methodical way and suggest methods

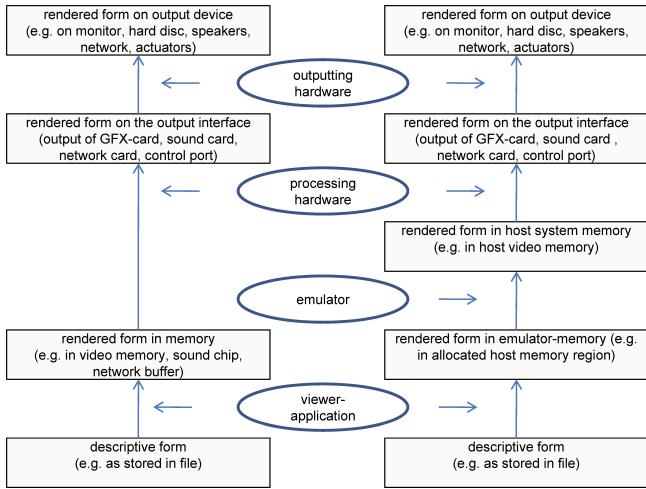


Figure 2: Different forms of a digital object in a system's memory. On the left the layers of an original system are shown, on the right the layers of the system hosting the emulator are shown.

to automate the process of evaluation to some extent. We described the different methods to automate input to the rendering environment to ensure that changes in manual handling of the digital object can be ruled out as a cause for changes in the rendering process. We also described the levels on which information can be extracted from the emulation environment as shown in Figure 2.

In this paper we show how we apply some of the concepts presented in the theoretical work on an existing emulator we presented in [7]. We implement automated input and extraction of data from the emulated environment. We identify key characteristics of the rendering process which can be measured automatically to not only evaluate the emulation environment but also to help improving the emulator.

3. VIDEOPIAC EMULATOR O2EM

The emulator we chose for implementing features for evaluation, O2EM², was previously described in [7]. It emulates a home-computer system introduced in 1978 as well as an updated version of the system released in 1983. The original system is both usable as a video game console by inserting cartridges with games, but due to its built-in keyboard it was also used as a home-computer with a special BASIC-cartridge. In this home-computer-mode the system was able to run BASIC programs and also load and save data to an external tape recorder.

In our previous work we implemented features in the emulator to make it usable for digital preservation purposes from a user's point of view (e.g. data exchange between the emulated and the host system). To actually be able to use an emulator in a digital archive, however, we need the possibility to evaluate the rendering process of digital objects more objectively and in an automated way. Based on our theoretical work in [8] we decided to implement the following

²O2EM Sourceforge Page - <http://o2em.sourceforge.net/>
O2EM DP version - <http://www.ifs.tuwien.ac.at/dp/o2em>

features:

Event-Log The original system can be controlled by using either the keyboard of the system or joysticks. In interactive applications (and especially video games) timeliness and type of input usually have a major influence on the behavior and thus resulting rendering of the digital object. Besides recording the points and type of input, we also wanted to log other events like file access (reading / writing to files in home-computer-mode) and the start of drawing an image frame (i.e. the start of the Vertical Blank period on the original system), to allow us to make statements about the correct timing of the emulator compared to the original system. Additionally, we recorded user-driven events in the emulator such as triggering a screenshot or a memory dump.

Automated Input The previously created event-log was defined in a form that is usable also as a command-file for the emulator, allowing us to automatically apply input to the system as well as create screenshots and memory dumps at specified times.

Memory Dumps We also implemented a feature to trigger memory dumps of the different memory regions in the system, including the hardware registers of the multimedia processors. This allows us to not only rely on screenshots of the emulator or files saved in the home-computer-mode as a way to extract data from the rendering process.

The next sections describe in detail the design decisions taken when implementing these features.

4. RECORDING OF EVENTS

The migration of an object lets us to some extent draw conclusions about the digital preservation action taken by comparing the object's properties before and after migration. Yet we need to draw conclusions on the rendering process of a digital object. We have to extract that information from the rendering environment and not from the object. To allow this, we need to implement an event-log of the rendering process in the rendering environment, e.g. an emulator or the new viewer application. We decided to include the following information in the log-file:

Executed Cycles The system emulated in the rendering environment usually runs at a different clock speed than the host system. Therefore we decided on the number of executed cycles as the main indicator of timing of when an event appears. This adds value for automated testing, as during an unsupervised test the emulator can be run without any speed limits, thus reducing the time needed for testing.

Elapsed Time As an additional time measurement we also record in the log-file the actual elapsed time since the rendering process was started. This measurement gives us an indication of how the emulator speed is perceived by a user of the emulator and may be used to normalize for timed events driven by a clock-based system rather than execution cycles based timing.

Drawn Frame As an additional timing measurement we record for every event in which 'frame' (unique consecutive image produced by the video hardware of the system) the

event was registered. (Note: For the purpose of this study we focus on the screen rendering for ease of presentation. Other forms of output rendering, such as acoustic behavior or output on other devices such as storage units, are considered in a similar manner.)

Recorded Event For each event we record the type of event as a code and as full text (for easier human readability).

Additional Infos Additional information on the recorded event is included, e.g. the key that has been pressed, the file that has been accessed etc.

To easily import the resulting file in spreadsheet applications for further processing and analysis we decided to use a comma separated value (CSV) format escaping commas that form part of the input in the log. When starting the emulator the event-log file that should be created can be specified as an extra parameter.

The following different types of events were defined for the system emulated in the emulator:

4.1 Controlling the Environment

To be able to evaluate the rendering process reliably, we have to make sure that the rendering is always exactly the same under the same conditions applied to the rendering environment, i.e. the emulator is deterministic in its behavior. Lamport et al. describe deterministic algorithms as „algorithms in which the actions of each process are uniquely determined by its local knowledge“ ([11]). This means that for any object rendered in the environment relying on external input to the rendering environment (e.g. user input, network activity, access to files on the host system) the type of input as well as the actual input data have to be stored to be able to provide the same data on a re-run for evaluation purposes.

The emulator O2EM (and the original system it emulates) supports user input in the form of key presses and joystick input. The hook-point for recording these events for the event-log is the interface in the emulator between the emulated environment and the host environment, i.e. when the emulator detects that the emulated process is trying to access the hardware registers that usually store the input values and provides the host system input instead. By recording the exact cycles already executed in the rendering when accessing this information, we are able to provide the same information when re-running the rendering process.

Reading files in home-computer-mode as a different type of providing external data to the rendering environment was recorded in the event-log, to let the digital archivist know that for later evaluation of the emulator these files have to be present besides the actual digital object, as they also potentially influence the rendering process.

4.2 Extraction of Data

As a basis for comparing the results of the emulation process, it is necessary to extract not only events but actual data as a result of the rendering. In Figure 2 we show different levels on which a rendered object exists during the rendering process. From inside the emulator we have access

to two different forms of rendered information: the form in the (emulated) memory of the system (e.g. hardware registers of the multimedia processor, usually triggering an output on the original system) as well as the form that is already translated to the host system (e.g. a rendered screen based on hardware registers of the emulated system's video hardware).

In O2EM a feature to save screenshots of the currently displayed image was already present. We enhanced this feature to create an event-log entry including (as every log entry) the executed cycles up until the point in the rendering the screenshot was taken. Additionally, we implemented a feature that works similar to saving screenshots that lets the user save the different emulated memory regions of the host system: memory internal to the processor, main system memory external to the processor, multimedia hardware registers memory and, if available, the emulated home-computer-mode memory. Additionally, in home-computer-mode files can be stored externally, which also influences the rendering process. The process of writing these files was also recorded in the event-log.

Under the assumption that the emulator works as a deterministic process, extracting data under the same external conditions (e.g. the exact same input applied) at the same point in the rendering process should provide the exact same result files.

4.3 Additional Events

In addition to the events described above, we also defined two other special event types for the log:

Vertical Blank The vertical blank is the period before the drawing of a new frame is started. It was an important event used to synchronize events on the screen to a fixed timing. We implemented this event to let us draw additional conclusions about how the number of cycles executed and the frames being drawn relates to the original system's timing.

Emulation Start For O2EM we record information about the cartridge image file that was rendered (filename and a checksum), as well as name and version number of the emulator and the date and time the log was created. This metadata gives us additional information about the rendering process for which the log was recorded.

Emulation Stop The information that the rendering process was stopped, the total number of cycles executed, the number of frames drawn and the elapsed time is recorded in the event-log.

5. AUTOMATED EXECUTION

Recording the events of a rendering process is only the first step in validation and verification of the digital preservation action. Especially if the rendering environment changes between execution of the digital preservation action and the re-deployment of the digital object at a later point in time, it is necessary to verify the correct rendering of the object in the new environment.

To be able to compare the rendering between validation (the time the digital preservation action was initially performed)

and verification we need to make sure that the external conditions influencing the execution are unchanged. This means that any manual input or external data applied to the rendering environment has to be the same as when the preservation action was initially validated. By recording these external events in a rendering environment and applying them at a later point in time to the new environment, we can compare the outcome of the rendering process.

In the emulator O2EM we implemented a feature to use the earlier described event-logs as command files. All external events and triggered data export actions recorded in the event-log file are automatically provided to the emulator using the command file. Actions are read from the command file and applied to the emulator when the specified number of cycles have been executed. In a deterministic emulator this means that the relevant actions are applied at the same time in the rendering process as they initially had been recorded.

In detail the following actions were implemented:

Keyboard and Joystick Input The manually recorded input events are applied at the exact same cycle count as initially recorded. The action from the command file is (similarly to the recording of the input for the event-log) interpreted once the emulator invokes the interface in which the emulated system tries to receive input from the host system. In a deterministic emulator the number of cycles executed until this check is performed does not change between renderings of the same digital object.

Screenshot and Memory Data Extraction The manually triggered extraction of data that has been recorded in the event-log file is automatically executed once the executed cycles stated in the command file are reached. Additional extractions can be inserted manually. This way it is possible to extract both a screenshot and all memory regions at the same point in the rendering process.

End Emulation The initial event-log record of the emulation stop also stops the emulation in the re-run once the action is encountered in the command file. This allows for automated and unattended testing of the emulator.

By first recording external events and later applying the event-log as a command file for a new version of the emulator (or even a different emulator) it is possible to automatically test an emulator. If the resulting data extracted at significant points in the rendering process is identical, we have a strong indication that the rendering process is unchanged.

6. KEY CHARACTERISTICS OF RENDERING PROCESS

Analyzing the event-log and using the features implemented in the emulator, we identified meaningful key characteristics of the rendering process, to see how the logs can help us evaluate if the rendering stays true to the original system or how it differs between different emulators (or different versions of the same emulator).

Deterministic Rendering The most important characteristic of a rendering environment is that the rendering process must be deterministic. This means that the emulator has to

perform the same rendering process under the same inputs. This is of crucial importance to the evaluation, as only a deterministic process lets us compare different renderings of the same object and the results of it.

Cycles Executed vs. Emulation Time Another characteristic we can extract from the rendering log is how many CPU cycles have been executed during the course of the emulation. If we compare these with the cycles that would have been executed on the original system (using the known clock rate of the original system), we can calculate the deviation in speed of the rendering process compared to the original system.

Executed Cycles per Frame By measuring the cycles that are executed per frame, we can see if the timing is correct. As we know the clock rate and the number of frames drawn on the original system from the systems specifications, we can evaluate any discrepancies to the original hardware.

Time Needed to Draw a Frame By evaluating the time that is needed to draw a frame and knowing how many frames are drawn per second (and thus the time the drawing of one frame should take) this characteristic also helps us evaluating the timing of the emulator.

Frames per Second Determining the frames per second we can see if the emulator is running slower than the original system is supposed to. If the emulator is in fact not fast enough, we can see from the event-log which of the drawn frames took too long to calculate and what external events happened during the slow frames.

Accessed External Sources By implementing a log for all interfaces between the emulated and the host environment, we also know which external resources (files, network, etc.) are used by a digital object. By logging the data that is transferred, we can decouple and simulate external interfaces at a re-run of the rendering process.

Using these key characteristics, we can evaluate an emulator, but also draw conclusions on the rendering process - not only in general for the rendering environment, but for specific digital objects. Re-running the same automated test in the emulator we can evaluate if the emulator works deterministic. Re-running the automated test of a deterministic emulator on a new version of the emulator we can test if the emulator still works correctly. Finally re-running the test in a different emulator for the same system, we can compare the results of these emulators.

7. EVALUATION OF O2EM

In this section we describe some of the experiments we performed on different digital objects suitable for the emulator we adapted. We describe the steps undertaken and the results of the rendering processes as well as the analysis of the resulting event-log files.

7.1 Video Game: Terrahawks

As a first digital object we chose a video game running in the standard mode of the emulator emulating a Philips Videopac G7000 running in European timing mode (PAL video stan-

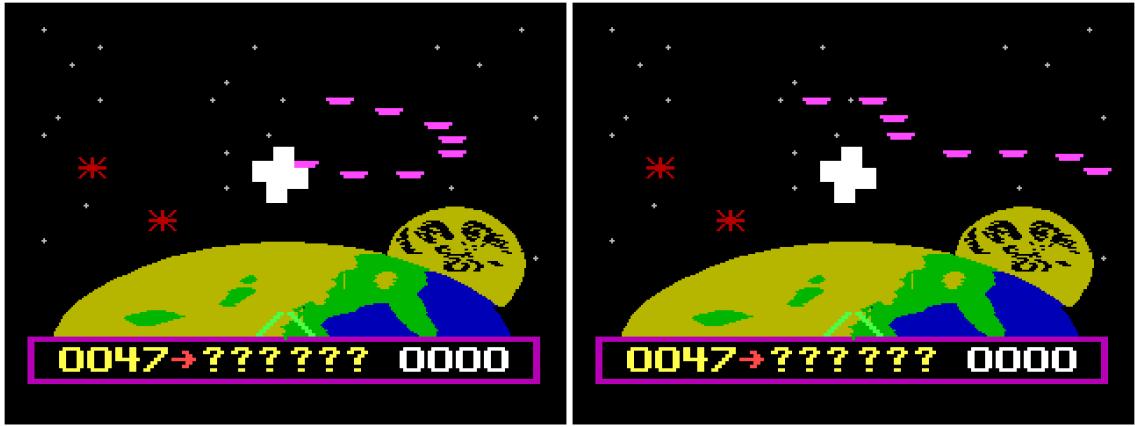


Figure 3: Non-deterministic rendering of *Terrahawks* - result of initial recording on the left, re-run on the right.

dard). We chose a video game as one of the objects because those are usually the most timing sensitive objects on the chosen hardware.

We chose the game *Terrahawks* that creates a random impression using external events to change the gameplay on every execution, to see if repeated execution of the game will produce the same rendering results, i.e. if the rendering process can be made deterministic.

As first step the emulator was started and one gameplay was recorded, both input using joystick but also some key presses (to enter letters for the highest score). A screenshot was taken after the game resulted in the player loosing his life (at which point the game just restarts, showing the new highest score on the bottom). In a second step the emulator was restarted with the event-log file given as a command-file. The previously recorded input was applied automatically during the rendering process. However the resulting screenshot taken at the same point in the rendering process as the original screenshot differed from the initial run of the emulator as shown in Figure 3.

A closer look on the emulator source code revealed that the emulation process was not entirely deterministic (i.e. independent from external factors), as the emulation of one of the hardware components, a voice synthesis module, was actually simulated using sound samples. The status check in the emulated code of this component was connected to the actual completion of playing the sample of the host system, an event the emulated environment had no control over. By deactivating the voice component, the emulation process was made deterministic and when the experiment was repeated, the results were identical on each re-run.

As timing in video games (especially action games) is crucial for the game experience, we used the rendering log to compare the timing of the real hardware (known due to the original system's schematics) to the values measured in the log as described in Section 6. The measured values as well as the expected values calculated from the original system's specification can be seen in Table 1.

Characteristic	Calculated	Measured
executed cycles per frame	7882	7259
executed cycles per second	394100	435540
frames per second	50	60
seconds per frame	0,02	0,0165

Table 1: Calculated versus measured key characteristics taken from the event-log of running *Terrahawks* in O2EM.

Based on these results it can be seen that due to the evaluation log we detected another error in the emulator. Even though the emulator was executed with the timing set to European TV-standard PAL timing (50 frames per second), the emulator was still rendering 60 frames per second as in the North American TV standard NTSC. The time taken for each frame was consistently 1/60 of a second, which is correct based on NTSC timing. The emulator was running fast enough to render every frame in less time than the original system would have needed, keeping the subjective feeling of speed for the user steady. Furthermore, it can be seen in Table 1 that the timing inside the emulator is not cycle-correct, thus timing-sensitive applications would not run correctly.

The findings about the incorrect timing were used to fix the errors in O2EM and improve the timing in the emulator, thus helping us to get a better rendering environment.

7.2 Application: Cassa

As a second example we chose not a video game but an application that runs in the home-computer mode of the system. We chose an application that allowed us to save data to the external tape drive and reload the data and render it during later use. The application was a BASIC-program distributed with the system in Italy, allowing the user to keep track of income/spendings per month over a year. We started the computer in home-computer mode, loaded the program, entered various fictitious data and saved the data in the program. For the actual evaluation we recorded the following process in the event-log: starting up the emulator in home-computer mode, loading the program (of which we

Characteristic	limited	no limit
total executed cycles	49201503	49201503
total frames drawn	6778	6778
total emulation time	136.426	10.512

Table 2: Characteristics for testing the application Cassa with original (=limited) and unlimited speed.

took a screenshot as seen on the left in Figure 4), loading the data into the program and displaying the data as also shown on the right in Figure 4. So not only the recorded user input but actual data loaded from an external drive influenced the rendering (i.e. what was shown on the screen).

To test the emulator in home-computer mode for determinism, we not only recorded screenshots (as due to the missing random element in the application those would most probably be similar), but also save the memory content of all different memory regions along with the screenshot of the displayed data (i.e. after the data was loaded into memory). We ran the test under two different settings in the emulator, first with speed limited as a user would usually experience it, and a second time without speed limit, simulating a verification where the test should be performed as fast as possible. We compared all the exported data files (screenshot and memory) with the result, that in all cases the files were exactly the same. As for the timing of the different runs as shown in Table 2, we can see that on our system the unlimited test executed the exact same test in only 7.7% of the time needed for a correctly timed emulation while creating the same results.

In the event-log we could also see the external files that had been loaded. These included not only the application Cassa itself, but also the file used for storing user entered data. In a real-life scenario this would enable us to identify which resources had been accessed and keep (or simulate) the necessary data for a later verification of the rendering of the preserved application.

8. CONCLUSIONS AND FUTURE WORK

In this paper we presented how previous conceptual work on evaluating emulators can be applied to evaluate the rendering process of different types of digital objects in an existing emulator. We first introduced the event-log of the rendering process with different properties that allow us to re-run a rendering in the same environment and potentially also in different ones. We showed the different kinds of events that have to be recorded depending on the original system. The different types of external data that can influence the rendering process have been explained as well as the different types of data that can be exported from the rendering environment for a comparison of different rendering processes. We then explained how the event-log can be used to automate the process of applying the same input data to the emulator to ensure a deterministic rendering of the digital object. After introducing the different key characteristics of the rendering process we identified in the event-log, we evaluated two different digital objects in the emulator O2EM and explained how the event-logs helped us to identify flaws in the rendering process.

Theoretical work we presented in [8] was successfully implemented in the emulator O2EM.

We rendered different objects in the emulator and analyzed the event-log files, which led us to the following conclusions:

Deterministic Emulation Automatically evaluating emulators by comparing the rendering results at different points in the rendering requires that the rendering environment behaves the same provided with the same external data. In the case of the game 'Terrahawks' evaluated in Section 7.1 the emulation was initially not deterministic, leading to different results of the rendering process, even though the obvious external data (user input) was kept constant. Only by making the rendering process deterministic, we could successfully compare the renderings in consecutive runnings of the emulator. This would also be the basis for later comparison of the rendering to later emulator versions or even other emulators.

External Data The external data needed to create a deterministic rendering is the one that is passed up from the host environment into the emulated environment. By recording the data that is transferred on these interfaces, we can apply the same data at the same point in the rendering process at a later time ensuring a deterministic rendering process. With the application 'Cassa' we showed that the external events (file access and user input) can be tracked in the event-log. External resources can then either be stored for a re-run for validation purposes or even simulated if the resources are no longer available (e.g. an external Web services).

Key Characteristics Using the key characteristics about the rendering process which we extracted from the event-log we were able to draw conclusions on the correctness of the emulation process. Especially deviations in handling the timing in the emulator were detected, assisting the emulator authors in improving the rendering process. Obviously when extending the described characteristics to more complex systems, additional characteristics could be found. Additionally to the time needed to draw a frame on the screen, similar measures could be captured for other output devices, e.g. port communications etc., where the timing of events needs to be captured, normalized and compared.

Automation of Evaluation Applying the external data to the rendering process not only gives us a possibility of creating a deterministic rendering, we can also automate the process of evaluating a rendering environment by applying the user input to a digital object automatically. This way interactive digital objects could be tested automatically on re-deployment in a new environment to see if the rendering is the same as at the time they have been preserved. We also showed that for this automated evaluation we not necessarily have to run the rendering process at the original system's speed, as all the automation is based not on time passed but on CPU cycles executed in the rendering environment, thus massively speeding up the process of the validation.

Overall we successfully implemented some of the concepts described in [8] in the existing emulator O2EM. This not only allowed improving the emulator for more accuracy, but also gave us a better understanding of the evaluation of ren-



Figure 4: Two screenshots taken during the rendering of Cassa - before starting the loading process on the left and after displaying data on the right.

dering environments in general. We showed that it is possible to automate the process of evaluating interactive objects beyond the manual testing of emulators with human interaction.

Future work has to be done on applying the concepts to more complex rendering environments like virtual machines that are more interwoven with the host system. Input and output events would have to be defined for more complex systems to catch all the events that are needed to make the rendering environments deterministic.

9. ACKNOWLEDGMENTS

This research was co-funded by COMET K1, FFG - Austrian Research Promotion Agency and by the European Commission under the IST Programme of the 7th FP for RTD - Project ICT-269940/TIMBUS.

10. REFERENCES

- [1] J. Barateiro, D. Draws, M. Neumann, and S. Strodl. Digital preservation challenges on software life cycle. In *16th European Conf. on Software Maintenance and Reengineering (CSMR2012)*, 3 2012.
- [2] C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman. Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries*, 10(4):133–157, 2009.
- [3] C. Becker, H. Kulovits, A. Rauber, and H. Hofman. Plato: a service-oriented decision support system for preservation planning. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL'08)*. ACM, June 2008.
- [4] C. Becker, A. Rauber, V. Heydecker, J. Schnasse, and M. Thaller. A generic XML language for characterising objects to support digital preservation. In *Proc. 23rd Annual ACM Symposium on Applied Computing (SAC'08)*, volume 1, pages 402–406, Fortaleza, Brazil, March 16-20 2008. ACM.
- [5] A. Brown. Automatic format identification using PRONOM and DROID. *Digital Preservation Technical Paper 1*, 2008. http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/automatic_format_identification.pdf.
- [6] M. Guttenbrunner, C. Becker, and A. Rauber. Keeping the game alive: Evaluating strategies for the preservation of console video games. *International Journal of Digital Curation (IJDC)*, 5(1):64–90, 2010.
- [7] M. Guttenbrunner and A. Rauber. Design decisions in emulator construction: A case study on home computer software preservation. In *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPres 2011)*, pages 171–180, 11 2011. Vortrag: iPres 2011 - 8th International Conference on Preservation of Digital Objects.
- [8] M. Guttenbrunner and A. Rauber. A measurement framework for evaluating emulators for digital preservation. *ACM Transactions on Information Systems (TOIS)*, 30(2), 2012.
- [9] M. Guttenbrunner, J. Wieners, A. Rauber, and M. Thaller. Same same but different - comparing rendering environments for interactive digital objects. In M. Ioannides, D. W. Fellner, A. Georgopoulos, and D. G. Hadjimitsis, editors, *EuroMed*, volume 6436 of *Lecture Notes in Computer Science*, pages 140–152. Springer, 2010.
- [10] M. Hedstrom, C. Lee, J. Olson, and C. Lampe. The old version flickers more: Digital preservation from the user's perspective. *American Archivist*, 69:28, 2006.
- [11] L. Lamport and N. Lynch. *Handbook of Theoretical Computer Science, Volume B: Formal Models and Semantics*, chapter 18, pages 1157–1200. Elsevier Science Publishers B.V., 1990.
- [12] M. Thaller. Interaction testing benchmark deliverable PC/2 - D6. *Internal Deliverable, EU Project Planets*, 2008. http://planetarium.hki.uni-koeln.de/planets_cms/sites/default/files/PC2D15_CIM.pdf.

Digital Preservation of Newspapers: Findings of the *Chronicles in Preservation* Project

Katherine Skinner Educopia Institute 1230 Peachtree St., Su 1900 Atlanta, GA 30309 404-783-2534 katherine@metaarchive.org	Matt Schultz MetaArchive Cooperative 1230 Peachtree St., Su 1900 Atlanta, GA 30309 616-566-3204 matt.schultz@metaarchive.org	Martin Halbert University of North Texas 1155 Union Circle #305190 Denton, TX, 76203 940-565-3025 martin.halbert@unt.edu	Mark Phillips University of North Texas 1155 Union Circle #305190 Denton, TX, 76203 940-565-2415 mark.phillips@unt.edu
--	---	---	---

ABSTRACT

In this paper, we describe research led by Educopia Institute regarding the preservation needs for digitized and born-digital newspapers. The *Chronicles in Preservation* project, builds upon previous efforts (e.g. the U.S. National Digital Newspaper Program) to look more broadly at the needs of digital newspapers in all of their diverse and challenging forms. This paper conveys the findings of the first research phase, including substantive survey results regarding digital newspaper curation practices.

Categories and Subject Descriptors

E.1 [Data Structures]: *distributed data structures*. H.3.2 [Digital Libraries]: *Information Storage, file organization*. H.3.4 [Systems and Software]: *distributed systems*. H.3.6 [Library Automation]: *large text archives*. H.3.7 [Digital Libraries]: *collection, dissemination, standards, systems issues*.

General Terms

Management, Documentation, Performance, Design, Reliability, Standardization, Languages, Theory, Legal Aspects, Verification.

Keywords

Archival Information Packages, Data Management, Digital Archives, Digital Curation, Digital Libraries, Digital Newspapers, Digital Objects, Digital Preservation, Distributed Digital Preservation, Ingest, Interoperability, Micro-Services, Repository Software, Submission Information Packages.

1. INTRODUCTION

U.S. libraries and archives have digitized newspapers since the mid-1990s using highly diverse and ever-evolving encoding practices, metadata schemas, formats, and file structures. Increasingly, they are also acquiring born-digital newspapers in an array of non-standardized formats, including websites, production masters, and e-prints. This content genre is of great value to scholars and researchers, and it is in critical need of preservation attention. The diversity of file types, formats, metadata, and structures that constitute this genre raises two major concerns: How can curators ready these collections for preservation? How may they conduct efficient repository-to-repository transfers from their local systems into digital preservation repositories?

The US National Endowment for the Humanities (NEH)-sponsored “Chronicles in Preservation” project is enabling the Educopia Institute, in collaboration with the MetaArchive

Cooperative, the San Diego Supercomputer Center, and the libraries of University of North Texas, Penn State, Virginia Tech, University of Utah, Georgia Tech, Boston College, Clemson University, and the University of Kentucky, to investigate these issues through the following research questions:

1. **How can curators effectively and efficiently prepare their current digitized and born-digital newspaper collections for preservation?** We are documenting guidelines and available tools for the evaluation and preparation of a diverse set of newspaper collections for preservation. We are analyzing the costs and benefits of data preparation and studying how best to lower obstacles to preservation.
2. **How can curators ingest preservation-ready newspaper content into existing digital preservation solutions?** The project team is studying existing mechanisms for repository exchange. We are building software bridges to facilitate the exchange of newspaper collections between partners’ local repository systems and distributed digital preservation (DDP) frameworks

This paper conveys the findings of the first phase of our project work, including substantive survey results we have gathered and analyzed regarding digital newspaper curation practices. In it, we begin by exploring the range of issues that born-digital and digitized newspaper content raises for curation and preservation practices. We then share information regarding our project findings and recommendations for near-future work.

2. THE CALF-PATH SYNDROME

*...A hundred thousand men were led
By one calf near three centuries dead.
They follow still his crooked way,
And lose one hundred years a day,
For thus such reverence is lent
To well-established precedent.*

-Sam Walter Foss, “The Calf-Path”

The story that the nineteenth century librarian and poet Sam Walter Foss tells in his poem entitled “The Calf-Path” is the story of a calf that perambulates through a wilderness, leaving behind a crooked trail that is gradually built up by subsequent animals and then humans. Over the course of a century the twisted trail becomes a road and eventually a highway through the center of a great metropolis. The poem is a humorous cautionary tale about the dangers of blindly following unexamined precedents.

The poem is a useful allegory concerning the problems that digitization and digital preservation programs may encounter when growing over time. Many such programs have humble origins in underfunded libraries and other cultural memory organizations, and are begun informally by a small number of staff who often “make it up as they go along.” As such programs blossom and achieve larger scale they often unwittingly preserve unexamined workflow precedents, much like the humans following the crooked trail of the calf in the poem. Often, these “calf-path” workflow problems are not evident to the individuals following the pre-established precedents. Rather, staff members are so busy trying to move more digital content through these well-established but inefficient practices that they never have the opportunity to step back and assess the overall efficacy of established workflows. The authors have examined the calf-path syndrome in digital preservation programs previously. [1] The calf-path syndrome is evident in most existing digital preservation programs for newspapers. We will occasionally invoke the calf-path syndrome in critiquing programs examined in this paper.

3. SIGNIFICANCE

The curation and long-term preservation of digital newspaper content presents unique challenges that are not fully understood and that demand additional research to ensure the survival of today’s digital newspaper collections for tomorrow’s researchers.

3.1 Newspapers as a Preservation Problem

Libraries and archives provide researchers with access to millions of digitized pages of historic newspapers. Some of these newspapers were scanned from print copies; others from microfilm. Some were digitized in-house; some outsourced to vendors. The scanning and encoding processes used in the digitization of historical newspapers vary wildly, as do the repository structures and storage media in which they are held.

Further complicating this digital genre, most newspaper producers shifted their operations to digital production by the beginning of this century. Increasingly, these born-digital print-production files are being acquired by libraries and archives. Many news groups also maintain websites that include non-AP wire materials of great value to researchers. As with digitized newspaper files, these born-digital files represent a range of format types (including websites, production masters, and e-prints) and are arranged in a wide variety of file structures and repository systems.

Digital newspaper files, then, are of increasing cultural and historical importance to researchers served by libraries, archives, and other memory organizations. One quality shared by nearly all of these diverse digital newspaper collections is that they are not yet preserved. [2] The lack of standard or normalized practices for the curation of these digital newspaper collections both within individual institutions (where practices have changed over time and remediation of earlier collections has not been pursued) and across the nation makes digital newspaper collections a high-risk genre of content that presents significant preservation challenges.

Research has demonstrated clearly that content preparation and ingest are the most time-consuming and costly parts of preservation (creating SIPs and AIPs, in OAIS terminology). [3] The steps involved in preparing content include properly documenting a collection (ascribing descriptive, technical, and structural metadata to files and collections), ensuring its current and future viability (establishing that the files will render on

current and future media), and organizing the files so that they can be managed over time (attending to file naming conventions and file structures such as folder and sub-folder designations).

The more normalized a collection is, the easier (and thus less time intensive and expensive) the process becomes of creating SIPs and, upon ingest, AIPs. In the case of digital newspapers, our research demonstrates that news content held within one institution is likely to include multiple digitized collections with different encoding levels, metadata treatment, file naming conventions, file types, and file structures because these collections were digitized at different times according to different standards, often by different teams (including external vendors). Also, these collections often are held in different repository systems.

For those institutions that are collecting born-digital newspapers, there are additional “calf-path” concerns. These collections are acquired in a wide range of ways, from hard-drive hand-offs of the master print-ready PDFs to Web crawls conducted upon newspaper Web sites. Because publishers vary widely in their own practices, the file types and file structures in these collections also include much variability. According to such factors, each of an institution’s digital newspaper collections may need individualized analysis to ready it for ingest into a preservation environment.

Unsurprisingly, curators cite grave concerns about how they will be able to prepare such problematic collections for preservation, both from practical and fiscal perspectives. [4] With limited resources, how can institutions prepare their content for preservation, and how much data preparation is “enough” to suffice? To address this question, our research team has explored the applicability of the NDNP’s existing set of recommendations for digitization efforts to the diverse body of legacy and born-digital newspaper content curated by libraries and archives.

3.2 NDNP Standards

The goal of the NEH and Library of Congress-supported National Digital Newspaper Program (NDNP) has been to develop an Internet-based, searchable database of U.S. newspapers that explicitly addresses the long-term content management and preservation needs of these collections.

The foremost set of technical parameters defined by the program relates specifically to scanning resolutions and establishing standard, high-quality file formats for NDNP digitization (TIFF 6.0). The majority of the additional technical parameters developed by the program seek to establish quality requirements for uniform metadata (CONSER-derived), encoding levels (METS/ALTO), and derivative file formats (JPEG2000 and PDF w/Hidden Text). Each of these requirements is in keeping with current high standards for archival-quality digitization for image-based items, and prepares the collections for successful repository management as defined by the OAIS Model. [5] The NDNP, then, is establishing best practices with implications far beyond the “Chronicling America” collection. Other institutions that are beginning or continuing digitization of newspapers benefit greatly from these standards, which help to ensure standard levels of encoding, file types, and uniform metadata that are geared for inter-repository sharing and long-term data management.

However, a wealth of digitized and born-digital newspaper collections exists in libraries, archives and other institutions that has been produced and obtained over the past two decades in a broad range of format types. [6] These “calf-path” collections

have been encoded at varied levels, use a diverse array of metadata schemas, and are arranged in highly irregular file structures and repository systems. The NNDNP technical guidelines do not currently provide explicit recommendations for readying such “legacy” and born-digital collections for preservation.

Our research explicitly seeks to fill this gap, building on the stable foundation of the NNDNP guidelines to address additional content within the broader “newspaper” genre. Rather than taking a “one-size-should-fit-all” approach, we differentiate between two tiers of preservation preparation: the *essential* and the *optimal*. If data preparation guidelines aim only for the “optimal,” curators at institutions with limited resources will be unable to implement them. This would be detrimental to our main goal, which is to enable curators at institutions with a wide range of resources and collection types to begin preserving their digital newspaper collections. We seek to ensure that guidelines enable curators of various resource levels to preserve collections (again, defined as “ensuring that they may be accessed for as long as they are needed”), and that the standards and guidelines for the field do not themselves become preservation obstacles by making overly high demands that curators lack the resources to implement.

4. WHY DDP?

Recent studies and national initiatives (i.e., US NDIIPP) have urged the digital library community to explore collaborative technical and organizational solutions to “help spread the burden of preservation, create economies of scale needed to support it, and mitigate the risks of data loss.” [7] The library community has concluded “the task of preserving our digital heritage for future generations far exceeds the capacity of any government or institution. Responsibility must be distributed across a number of stewardship organizations running heterogeneous and geographically dispersed digital preservation repositories.” [8] Some early answers to this call embed collaborative practices in their technical and organizational infrastructures. For example, in distributed preservation repositories (e.g. Chronopolis, MetaArchive, CLOCKSS, Data-PASS), preservation activities occur within a dispersed network environment that is administered by multiple institutions. This approach combines geographic distribution with strong security of individual caches to create secure networks in which preservation activities may take place.

Such *Distributed Digital Preservation* (DDP) networks leverage inter-institutional commitments and infrastructures to support the requisite server infrastructures and to conduct necessary preservation activities in a local manner. In so doing, they capitalize on the existing infrastructures of libraries and archives (and in some cases, their parent institutions), simultaneously reducing costs and ensuring that digital preservation expertise is community-sourced, or built within the cultural memory community, not outsourced to third-party service providers.

Though the digital medium is relatively new, the conceptual approach taken by DDP practitioners is not. In the scribal era, this combination of approaches—geographic dispersal of content and secure storage environments—maximized the survivability of content over millennia. [9] Secure distribution helps content to withstand large-scale disasters (e.g., wars, hurricanes power grid failures) and more isolated, local-level events (e.g., media failures, human errors, hacking, fires).

In the last decade, many programs have developed using collaborative and distributed methodologies, and still others are in

pilot phases of their research and development work. Examples of proven approaches include MetaArchive (Private LOCKSS Network (PLN)), Chronopolis (SDSC’s iRODS-based service), and the Data-PASS Network (ICPSR/Roper Institute/Odum Institute partnership to preserve social science datasets using a PLN). Other experimental approaches show great promise, including Digital Preservation Network (DPN, bridging heterogeneous preservation environments), DuraCloud (DuraSpace’s cloud-storage-based environment) and LuKII (a German program that bridges LOCKSS’s cost-effective preservation with KOPAL’s usability and curation tools).

The demand for community-based initiatives hosted and managed by libraries and archives is strong. Surveys conducted by the MetaArchive Cooperative in 2009 and 2010 reveal that curators of digital newspaper content both need and actively seek implementable digital preservation solutions and models. Most institutions (80%) report that they do not aspire to build their own preservation repository due to the expense, technical expertise, and infrastructure required. Fully 73% of 2009 and 2010 respondents reported that they were interested in using community-based preservation networks, while only 30% reported interest in third-party vendor solutions. [10]

The Chronicles research project focuses on three approaches to preservation—MetaArchive, Chronopolis, and CODA—which share certain common characteristics, but use very different technologies to accomplish their goals. The three most salient similarities between these approaches are 1) they all use open-source technologies; 2) these are library-run, community-sourced ventures; and 3) these are *Distributed Digital Preservation* (DDP) approaches. Each of these approaches varies in other key areas such as ingest mechanisms, data management practices, organizational model, and recovery options.

4.1 MetaArchive Cooperative

The MetaArchive Cooperative is a community-sourcing network that preserves digital collections for more than 50 member libraries, archives, and other digital memory organizations in four countries. The Cooperative was founded in 2003-2004 to develop a collaborative digital preservation solution for special collections materials, including digitized and born digital collections. Working cooperatively with the Library of Congress through the NDIIPP Program, the founders sought to embed both the knowledge and the technical infrastructure of preservation within MetaArchive’s member institutions. They selected the LOCKSS software as a technical framework that matched the Cooperative’s principles, and built additional curatorial tools that layer with LOCKSS to promote the curation and preservation of digital special collections, including newspapers, Electronic Theses and Dissertations, photographs, audio, video, and datasets. In doing so, they created a secure, cost-effective repository solution that fosters ownership rather than outsourcing of this core library/archive mission. The Cooperative moved to an open membership model in 2007, and has expanded in five years from a small group of six southeastern academic libraries to an extended community of more than 50 international academic libraries, public libraries, archives, and research centers.

4.2 Chronopolis

The Chronopolis digital preservation network has the capacity to preserve hundreds of terabytes of digital data—data of any type or size, with minimal requirements on the data provider. Chronopolis comprises several partner organizations that provide a wide range

of services: San Diego Supercomputer Center (SDSC) at UC San Diego; UC San Diego Libraries (UCSDL); National Center for Atmospheric Research (NCAR); and University of Maryland Institute for Advanced Computer Studies (UMIACS). The project leverages high-speed networks, mass-scale storage capabilities, and the expertise of the partners in order to provide a geographically distributed, heterogeneous, and highly redundant archive system. It uses iRODS (Integrated Rule-Oriented Data System) to federate three partner sites and replicate data, BagIt to transfer data into the storage locations, and ACE (Audit Control Environment) to monitor content for integrity.

4.3 University of North Texas

The University of North Texas has constructed a robust and loosely integrated set of in-house archiving infrastructures to manage their digital collections, including a delivery system (Aubrey) and a Linux-based repository structure (CODA). The underlying file system organization of digital objects is tied to a UNT-specific data modeling process that relies on locally developed scripts and micro-services to generate and define all master, derivative, related objects, metadata, and other information that may be tied to a single digital object in order to effect archival management and access retrieval. This archival repository solution has been designed with open source software and relies on loosely bundled specifications to ensure on-going flexibility. UNT's archival repository implemented its integrated offsite replication in 2010. The micro-services that support the current instance of CODA are being experimented with for optimizing workflows across both instances of the repository.

5. SURVEYING DIGITAL NEWSPAPERS

The Chronicles in Preservation project has investigated a diverse array of digital newspaper content and its associated preservation needs across a broad stratum of institutions. This took the form of an extensive survey and set of interviews that were carried out beginning in October 2011. [11] The eight academic libraries participating in the project were asked for detailed information about the range of digital newspaper collections they curate (e.g., file formats, encoding practices, etc); the repository infrastructures they use to support this content; and their requirements for archival ingest and long-term distributed digital preservation. A summary of the survey findings follows.

5.1 Preservation Formats, OCR & Metadata.

The surveyed content curators cited divergent needs and practices regarding what image formats they produce, manage, and intend to preserve. Most surveyed libraries report using TIFF as their primary master image format (the exception, Virginia Tech, works exclusively with born-digital content—HTML and PDF). The respondents also reported using a range of derivative file types, including PDF (7 libraries), JPEG2000 (6 libraries), JPEG (3 libraries), xml (2 libraries), and HTML (1 library).

Preservation ambitions vary across the surveyed libraries. Some locations intend to preserve only their master TIFF images (Clemson, University of Kentucky, University of Utah, and UNT). Others also focused on their derivative JPEG and PDF images (Georgia Tech), and JPEG2000 images (Boston College). All respondent libraries report that no file format used in their newspaper curation practices has become obsolete to date. All likewise report that they have only normalized and migrated files for the purposes of producing derivatives for access. Four of the

respondent libraries report using JHOVE for file format identification and/or validation purposes.

In addition to the array of target image formats mentioned above, all of the content curators are creating & maintaining a range of OCR formats (XML, PDF, ABBYY, METS/ALTO, ALTO, PrimeOCR, etc.) and metadata (Fedora Core, METS, MIX, MODS, customized Dublin Core, etc.) formats. In some cases, the collection/object-to-metadata relationships remain somewhat opaque to the content curators due to their reliance upon their repository software for metadata creation and maintenance. In several other cases, content curators are making use of METS to encapsulate their digital objects and various associated metadata. In most cases, the content curators were confident that their metadata could be exported from their repository systems in some form of XML for external processing.

5.2 Repository Systems & Features.

Content curators are using a diverse array of repository software solutions to manage their digital newspaper collections. These include licensed open-source solutions such as Fedora (Clemson) & DSpace (GA Tech), as well as licensed proprietary solutions such as CONTENTdm (Penn State; University of Utah), Olive ActivePaper (Penn State) & Veridian (Boston College). Other implementations range from University of Kentucky's (UKY) and University North Texas's homegrown infrastructures modeled on a micro-services architecture, all the way to the use of simple web servers (Penn State; Virginia Tech). It should be noted that with the exception of UKY and UNT, none of the repository solutions indicated above are aiming to be fully supported preservation systems. The systems reported are generally prioritized to support access. Only Georgia Tech is storing their master TIFF images in their DSpace repository instance (with backup support on-location). In most cases, master TIFFs or JPEG2000s are typically stored and backed-up on on- or off-site SAN or tape systems.

In order to prepare the content stored in these access-oriented systems for ingest into preservation systems, SIPs may need to be staged externally. It should also be noted that some dependencies exist at the level of metadata and object/collection identifier creation and export, as these systems provide custom-built or proprietary modules with varying degrees of flexibility for open- and user-defined conventions. Export utilities and HTML/XML parsers may need to be identified or developed to support their harvest and retention at ingest.

5.3 Data Management Practices.

Collection and/or object identifier schemes for content curators' repository environments spanned a wide range of implementations. Most of these schemes employ user- or system-generated persistent identifiers (e.g., Fedora PID at Clemson, DSpace Handles at Georgia Tech; Veridian custom URLs at Boston College; NOID and CDL Identity Service at UKY; CONTENTdm Reference URLs at University of Utah; Coda ARKs at UNT). Only three of these content curators have developed formal digital object identifier schemes external to these repository systems (Boston College and UNT). Boston College uses a standard code for a newspaper title, a CCYYMMDD date, and 3-digit image/page sequence number (e.g., bcheights/1921/05/21/bcheights_19210521_001.jp2). UNT assigns a unique identifier at the digital object level according to CDL's ARK specification. UKY makes use of NOID in conjunction with a locally developed identifier scheme. All content curators have indicated that the retention of any collection

and/or object identifiers is crucial for recovering their current repository environments. However, this warrants further investigation into the ramifications of decisions regarding what forms of the content are preserved (e.g., preserving master images and not derivatives) as this may hinder the recovery of an access-based repository environment.

5.4 Collection Sizes, Growth Rates & Change.

Reported collection size aggregations follow a number of models—some by title, some by issue, others by originating institution. Some aggregations are no more than 60 megabytes, others can reach as much as seven terabytes. The majority of collection aggregations that were surveyed stay well below half a terabyte. Content curators are systematically acquiring and adding new digital newspaper content according to a variety of schedules. University of Utah, University of Kentucky, and University of North Texas reported the most dynamic rates of acquisition—20,000 pages per month, 20,000 pages per quarter, and 40,000 issues per year respectively. Penn State also reported a robust rate of acquisition at approximately 75,000 pages annually. The majority of content curators however have relatively static or only mildly growing digital newspaper collections. Georgia Tech reported ten issues of growth per month, and Clemson University only one or two titles per year. Boston College could only speculate on future growth with two potential titles under negotiation, and Virginia Tech suggesting no future growth.

Content curators were surveyed for any existing change management policies or practices in the midst of such rates of growth. This was intended to account for image or metadata files that may have undergone repair or refreshment—tracking or associating versions of files through identifier or naming conventions for example. This was also intended to account for any changes to underlying technical infrastructure supporting local archival management—perhaps recording technical and administrative metadata through METS or PREMIS. None of the content curators, with the exception of UNT, had formal change management policies or could clearly identify repository or other system features that were accomplishing version management. UNT has a robust set of data management workflows that account for all events that take place on a digital object (files and metadata). They are also moving towards establishing workflows that track changes to technical infrastructure (hardware refreshment, system updates, etc.). Knowing the state of such local policies and practices can help institutions understand the degree to which such meaningful preservation activities may need to be accommodated or similarly maintained external to the content curator.

5.5 Preservation Preparedness

As detailed above, content curators are currently managing a range of well-supported digital formats for their digital newspaper collections. In most cases, content has been digitized to high archival standards. Master images are in TIFF format, and derivative access copies are in high-resolution JPEGs, PDFs, or JPEG2000s. Exceptions to these standards include a small subset of very early versions of HTML-encoded websites, and lower-resolution PDF master images.

As previously mentioned, none of the content curators we surveyed have performed format migration or normalization for the purposes of preservation. Among the surveyed libraries, file format identification tools like JHOVE, JHOVE2 or DROID are in moderate use (4 of the 8 institutions). None of the surveyed

content curators currently subscribe to format registry services such as the Unified Digital Formats Registry (UDFR). With the exception of one content curator, the use of PREMIS is not yet routine or programmatic. However, as also noted above several content curators are gathering administrative, technical, structural, and provenance metadata for the digital objects that comprise their digital newspaper collections. In some cases this metadata is being systematically generated at ingest through the use of JHOVE, and other system utilities, and being related to corresponding digital objects through use of METS, MIX & MODS—which can be bridged to PREMIS. When asked about near- to long-term capacity for creating and managing preservation metadata most content curators stated a current lack of familiarity with PREMIS, but noted their awareness of it and their potential staff capacity for integrating PREMIS in their local workflows in the future.

Beginning in Fall 2012, the Chronicles in Preservation project will enter the Transition and Documentation Phases, in which project staff will document the necessary preservation readiness steps that the project partners need to apply to their own very diverse holdings—both digitized and born-digital—for the purposes of experimenting with more robust preservation. These individualized “preservation preparedness plans” will be derived from the more general *Guidelines to Digital Newspaper Preservation Readiness* that we are currently producing. Like the *Guidelines*, these preservation preparedness plans will seek to document preservation readiness strategies for each institutional partner along a spectrum of the *essential* to the *optimal*.

This “spectrum” approach enables the content curators at our partner institution sites (as with the larger field addressed in the *Guidelines*) to understand the acceptable range of activities they may undertake in their preservation readiness practices. By documenting the *essential* and the *optimal*, we invite and encourage institutions to engage responsibly with preservation at the level they can currently handle without delay. We also make it possible for those with lower resources to understand the difference between their current activities (*essential*) and those to which they should aspire in the future (*optimal*). The *essential* recommended readiness steps to be taken may be achieved even given the limited resources and expertise that are typically available to the average content curator. These are what we consider non-negotiable activities, because to neglect them would undermine the long-term preservation of their content. The *optimal* workflows will ensure the highest standards in long-term preservation for those that do have the resources to pursue them now, and they will provide those institutions that can only aspire to the “*essential*” level today with benchmarks for later success.

We believe that taking this flexible approach to documenting preservation measures for digital newspapers will enable content curators to understand what they can begin doing in the short-term in the absence of high levels of resources and expertise, and will provide them with a foundation for the “*optimal*” curation practices to enhance their preservation capacity going forward.

5.6 Preservation Pathways

Each of the project’s three DDP sites has its own unique mechanisms for handling ingest, packaging AIPs, and effecting long-term preservation. During the surveys, content curators were asked a series of questions about their experience concerning digital newspapers with the general types of ingest-related technologies that each of the preservation sites use (e.g., web harvesting mechanisms, use of the BagIt specification, and the use

of micro-services). Aside from Virginia Tech's previous development work to ingest digital newspaper content into MetaArchive, and UNT's use of BagIt and various micro-services, none of the respondents have pursued these technologies for managing their digital newspapers.

Similarly, but with a different emphasis, content curators were surveyed for their preferences for ingest strategies. Suggested options included shipping hard drives, performing server-to-server copies, performing BagIt based transfers, or triggering web harvests on staged content. Half of the content curators (4 of 8) indicated a strong preference for shipping their hard-drives to preservation sites or allowing a preservation site to perform remote copying of data from a secure server connection, and half also showed a preference for the use of BagIt. Web-crawl strategies fared somewhat lower in terms of preference, with only two content curators listing this strategy as a first option.

6. DIGITAL NEWSPAPER CASE STUDIES

Following the survey, we conducted in-depth interviews with our partners. Below, we share information from University of North Texas (UNT) and Virginia Tech drawn from the focused interviews we have conducted. The UNT case study provides one possible pathway for rectifying the calf-path syndrome by carefully balancing the needs associated with inherited precedents against local needs for achieving scale and efficiency. The Virginia Tech case study illuminates the kind of meandering workflows that can arise when a preservation program inherits content streams from many pre-existing sources.

6.1 University of North Texas Case Study

The University of North Texas Libraries (hereafter UNT) are actively involved in a number of newspaper digitization and preservation activities. Beginning in the same year as its first NDNP award, UNT developed a comprehensive program to identify, collect, digitize and preserve newspapers from around the state of Texas with a program called the Texas Digital Newspaper Program [12]. The team at UNT leveraged the technical specifications of the NDNP program in all but one area for use in non-NDNP newspaper digitization as well as identifying several new workflows for the acquisition and processing of born-digital print masters from publishers around the state. All digitized and born-digital newspaper content is added to The Portal to Texas History [13] for end user access and also to the UNT developed CODA preservation infrastructure for long-term storage and management. To date UNT has made freely available over 750,000 pages (95,000+ issues) from 409 different titles via The Portal to Texas History.

6.1.1 Standards and Workflow

The UNT workflow for newspaper digitization and born-digital processing is heavily influenced by the *NDNP Technical Guidelines and Specifications* [14] that is comprised of a number of technical sub-specifications, all of which are important when trying to organize a large-scale newspaper digitization program like the NEH NDNP program or UNT's Texas Digital Newspaper Program. UNT found that these specifications provided a good starting point for refining its internal workflows and standards.

Source Material Selection: The NDNP specification advises use of second-generation negative film on a silver halide substrate. The specification also allows use of born digital images or images scanned from paper. UNT found it very important to use second-generation negatives for the best results in the digitization

process. For titles only available in print format UNT contracted with vendors to microfilm the title before the digitization process. Born-digital files are also collected from a number of publishers around the state. Typically these are the production print masters sent to the printers that are then delivered to the UNT team. The goal in each content stream is to ensure that the highest quality, most complete version of the title is being used for later processing.

Scanning: The NDNP specification describes the resolution and color space that is optimal for scanning content: 300-400 DPI using 8 bit grayscale. UNT views this as a minimum resolution, whether the scanning is performed by outsourced services or internally within the UNT Libraries. Born-digital print masters are converted from their delivered formats (usually pdf) into 400dpi, 24bit JPEG images which are used for subsequent processing. The delivered pdf masters are retained and stored with the object in the final archival package ingested into the CODA repository.

File processing: UNT aligns with the NDNP specification with regard to processing on the master files created in the digitization process. Scanned images are de-skewing to within 3% skew and cropping with a slight edge around the physical piece of paper, not just the text on the page. Born digital items are left unaltered other than occasional 90-degree rotation to properly align the text.

OCR: UNT utilizes the ABBYY Recognition Server for the optical character recognition (OCR) process when items are digitized in-house. The ABBYY software is operated in a cluster configuration with six nodes (52 cores) dedicated to the OCR process. UNT has found this tool to provide an appropriate tradeoff between quality, convenience and costs of OCR.

Serializing a newspaper issue to files: The NDNP specification describes the use of the METS and ALTO specifications to represent a newspaper issue on a file system. This is an area that UNT begins to depart from the NDNP specifications to allow for integration into local systems. OCR files from the ABBYY Recognition Server are converted into several legacy formats for representing bounding box information and indexed text. The master ABBYY XML file is also saved with the output files for later processing if the need arises. All pages associated with an issue are placed in a folder named with the following convention, yyyyymmddee (y=year, m=month, d=day, e=edition). Descriptive metadata is collected for each issue and stored alongside the page images in the issue folder and is used at a later point in the ingest process. A future area of development is the conversion of the proprietary ABBYY format into the standard ALTO format used by our NDNP projects to allow for a greater use of ALTO enabled workflows and tools.

Derivatives: The NDNP specification calls for creating a JPEG2000 and PDF for each page of newspaper. UNT currently creates JPEG2000 derivatives on ingest into its Aubrey content delivery system. In addition to JPEG2000 files, traditional JPEG images are created in a number of resolutions such as square, thumbnail, medium and large to provide a variety of viewing strategies for end users. UNT also pre-tiles each image loaded into The Portal to Texas History with the Zoomify tile format and stores these tiles in WARC [15] files.

Ingest: The UNT Libraries' ingests all digitized and born-digital newspapers into a locally developed system called CODA, which provides archival file management for digital content under its management. Each item ingested is assigned a globally unique ARK identifier that is used to request the item from CODA.

Summary: The UNT internal workflow is heavily influenced by the NDNP technical specifications, which constitutes an excellent set of specifications for libraries and vendors to use in digitizing and delivering newspaper content. These specifications can be used as a starting point for developing local workflows that take into account new content acquisition strategies and formats not covered completely by the NDNP program. One key aspect missing in the NDNP specifications that might be useful to the newspaper digitization community is an extension to allow for article level data to be encoded into the METS/ALTO format.

6.1.2 Avoiding the Calf-Path

The UNT case study demonstrates ways of avoiding the calf path by carefully comparing and analyzing competing requirements that derive from external precedents and internal optimization needs. This is possible when setting up a new or relatively new program at scale, but may not be possible when a program has long-standing inherited precedents. It may be very difficult to get off the calf path in some situations, as the following case study from Virginia Tech illustrates.

6.2 Virginia Tech Case Study

The digital newspaper collections of Virginia Tech represent a diverse and un-normalized legacy of digital content. Within the *Chronicles in Preservation* project, Virginia Tech is a good case study in dilemmas associated with born-digital content, since the university has not engaged in digitization but has hosted born-digital newspaper content for almost two decades. Virginia Tech began accepting web pages and PDFs from various local, regional, international news agencies in 1992. More than 19 gigabytes of news content has now accumulated at the university, which was received directly from the publishers in digital formats.

In 1992, the Virginia Tech library began receiving online news feeds from the two major newspapers in Southwest Virginia, ultimately resulting in over 400,000 text files documenting life in this region. In 1994 the library began capturing the university's newspapers, and in 1997 international news began arriving in PDF format. The 2,600 PDF files collected provide a context for studying Lebanon, Iran, and France in the local languages—Arabic, Farsi, and French.

6.2.1 Problems with Metadata

Metadata was not systematically collected for this body of content for many years, since the Virginia Tech staff working on these projects was quite limited and in the early search engines of the 1990's ignored metadata. Staff members to create metadata were gradually added with the intent of implementing a better practice for organizing the digital content being gathered.

The first step taken was to begin adding very basic article-level info derived from the text files comprising individual newspaper articles. An example newspaper for which this practice was implemented is the *Roanoke Times*, which began including date, author, edition, location, and title information in the text file headers circa 1996. These metadata elements could be parsed and used for indexing, access, and organization purposes.

Various ad hoc parsing scripts were developed over time to extract metadata from the news content feeds received at Virginia Tech, and normalize this metadata into Dublin Core elements. This practice was fragile, however, and prone to malfunction if the format of the feeds changed over time. Virginia Tech is still

considering how to effectively automate the generation of metadata for these content feeds. This is an example of the most difficult kind of calf-path to escape, a long-standing set of uncontrollable external data feeds that cannot be remediated.

7. PRESERVING DIGITAL NEWSPAPERS

Though the range of content needs for the various digital newspaper holdings are highly diverse, even within a single curatorial location, the concept of "standardizing" requires us to pursue uniform approaches and recommendations, both broadly through the *Guidelines*, but also within the individualized "preservation readiness plans." This applies not only to such tasks as exporting and compiling metadata or forward migrating to de-facto standard OCR formats such as ALTO, but also attempting to achieve common packaging and ingest measures.

7.1 Repository-to-Repository Exchanges

Data exchange challenges are complex and as yet unresolved, both within and well beyond the library and archives communities. The most successful data exchange models address issues that arise in specific genres of content, from emergency alert systems (OASIS) to social science data sets (DDI). [16] Most data exchange models to date—including those created for newspapers—have been used primarily to address the integration and federation of content for access purposes. How might the genre of interest here—newspaper data—be exchanged for preservation purposes? The issues involved in data exchange in the preservation context are twofold, involving both data structures (the way that the collections' constituent parts are stored and how the repository system uses those stored components to assemble an access view) and repository system export and ingest options (ways of moving content in or out of repository environments). Libraries and archives, as mentioned above, use many different types of repository systems to store their digital newspaper content. Each of these repository systems has expectations about how data is structured. The mismatch of these expectations between repository systems makes it difficult to move collections from one system to another while maintaining each collection's integrity and set of relationships. [17]

We are currently studying existing specifications for transfer to assess their applicability to the genre of digital newspaper content, including UIUC's HandS Project, TIPR, and BagIt. [18] To date, much of the interoperability and exchange work between access-oriented repositories and preservation repositories for collaborative frameworks, like those chosen for evaluation in this project, have happened in one-off fashion. For example, the MetaArchive Cooperative has successfully exchanged content with Chronopolis, and has also ingested content from DSpace, CONTENTdm, Fedora, Digital Commons, and ETDb repositories by creating "plugins" specific to each content contributor's collections. Likewise, there have been projects that have explored the use of DSpace with SRB/iRODS and Fedora with iRODS. These have been largely geared toward addressing an individual institution's collections and have been mapped in a straightforward pathway from DSpace to iRODS and Fedora to iRODS. Such work may help individual institutions, but it does not efficiently streamline the ingest process in a way that is relevant to the larger digital library and archives community when preserving their content in various collaborative solutions.

7.2 Towards Interoperability Tools

We are currently documenting the complexities involved in streamlining such access-to-preservation repository exchanges. We are encountering a range of issues, exemplified here by our preliminary research. As detailed above, during these investigations a number of questions have arisen regarding compatibilities between partner institutions' collections and both the access-oriented systems and the preservation systems being evaluated. For example, what data management components must be implemented in the MetaArchive and Chronopolis environments to facilitate, create, and update the administrative, preservation, and technical metadata that accompanies a potential exchange profile? Is UNT-CODA's micro-services based approach for preparing SIPs to become AIPs extensible to the MetaArchive and Chronopolis environments and could this approach provide flexible alternatives to requiring well-formed and standardized exchange profiles? Conversely, how do the UNT workflows for enhancing SIPs through micro-services interact with exchange packages that already include this information (e.g., Penn State's NDNP collections)?

To study these and other issues, the project's technical team is analyzing the applicability of existing efforts to move content between systems for meeting our project goals. We are also experimenting with BagIt to determine whether that transfer mechanism will accommodate the full range of digital newspaper packaging requirements as documented in the *Guidelines* and "preservation readiness plans." In conjunction with our Chronicles Committee and Advisory Board, the project team is also studying the benefits of and barriers to implementing PREMIS and METS for our partners' collections and for these preservation environments. All of these findings will be documented in a white paper that will be released in early 2013 via the project site: <http://metaarchive.org/neh>.

8. CONCLUSIONS

The first phase of the project facilitated our understanding of the current practices and workflow needs of newspaper content curators. It also substantiated our theory that a single unified workflow is not an optimal approach for engaging institutions in the process of readying their content for preservation. To encourage broad participation, we should not seek to establish a single workflow or exchange mechanism for preparing a collection for ingest across all three preservation systems explored in this project. Rather, we will aim to reduce barriers by establishing a range of guidelines and workflows and by building systematic approaches for exchanging content between common access-oriented repositories and mature preservation solutions.

9. ACKNOWLEDGMENTS

We greatly appreciate the generous support of the National Endowment for the Humanities through Award PR-50134.

10. REFERENCES

- [1] Halbert, M., Skinner, K., and McMillan, G. 2009. "Avoiding the Calf-Path: Digital Preservation Readiness" *Archiving 2009 Proceedings*. pp. 86-91.
- [2] Skinner, K. and McMillan, G. 2009. "Surveys of Digital Preservation Practices and Priorities in Cultural Memory Organizations." NDIIPP Partners Meeting, Wash. DC., DOI= www.digitalpreservation.gov/meetings/documents/ndiipp09/NDIIPP_Partners_2009_finalRev2.ppt; Skinner, K. and McMillan, G. 2010. *Survey of Newspaper Curators*. Educopia Institute; Skinner, K. 2012. Survey on Digital Newspaper Archiving Practices. Educopia (*forthcoming*).
- [3] Beagrie, N., Lavoie, B., and Woppard, M. 2010. Keeping Research Data Safe 2 Final Report. JISC/OCLC. DOI= <http://www.jisc.ac.uk/publications/reports/2010/keepingresearchdatasafe2.aspx#downloads>.
- [4] Angevaare, I. 2009. Taking Care of Digital Collections and Data 'Curation' and Organisational Choices for Research Libraries. *Liber Quarterly*, 19:1. DOI= <http://liber.library.uu.nl/index.php/lq/article/view/7948>.
- [5] Library of Congress. 2009. NDNP: Technical Guidelines. http://www.loc.gov/ndnp/guidelines/archive/NDNP_201113TechNotes.pdf.
- [6] As the Library of Congress underscored in a Broad Agency Announcement (BAA) as a *Draft Statement of Objectives on Ingest for Digital Content* (June 2010): "Some digital content types have remained relatively stable in format over time (such as digital versions of academic journals), while others (such as digital versions of newspapers and other news sources) have become increasingly complex, evolving with the Internet environment.... Some digital content types are relatively self-contained... while others ...contain (and/or are linked to) multiple digital content objects."
- [7] Fran Berman and Brian Schottlaender, "The Need for Formalized Trust in Digital Repository Collaborative Infrastructure." *NSF/JISC Workshop*, April 16, 2007: http://www.sis.pitt.edu/~repwkshop/papers/berman_schottlaender.html (last accessed 06/07/2012); Please also see the following reports: American Council of Learned Societies. (2006) "Our Cultural Commonwealth: The Report of the ACLS Commission on Cyberinfrastructure for the Humanities and Social Sciences" *American Council of Learned Societies*: <http://www.acls.org/cyberinfrastructure/ourculturalcommonwealth.pdf> (last accessed 06/07/2012); Blue Ribbon Task Force on Sustainable Digital Preservation and Access. "Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information" February, 2010. Available at: http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf (last accessed 06/07/2012), and the JISC/OCLC "Keeping Research Data Safe 2 Final Report" (previously cited), which have pointed to the economic challenges inherent in "silo"-based development and maintenance in the area of preservation.
- [8] Priscilla Caplan, "IMLS Funds TIPR Demonstration Project." *Digital Preservation Matters*, 2008. Available at: <http://preservationmatters.blogspot.com/2008/09/imls-funds-tipr-demonstration-project.html> (last accessed 06/07/2012).
- [9] Katherine Skinner and Matt Schultz, Eds., *A Guide to Distributed Digital Preservation*, Educopia, 2010. Available: http://www.metaarchive.org/sites/default/files/GDDP_Educopia.pdf (last accessed 06/07/2012).
- [10] For more on the surveys, please see Skinner and McMillan.
- [11] Schultz, M., Skinner, K., 2011. Chronicles in Preservation: Collections Assessment Survey. Educopia Institute.
- [12] The Texas Digital Newspaper Program (TDNP). Available at: <http://tdnp.unt.edu>

- [13] The Portal to Texas History. Available at:
<http://texashistory.unt.edu>
- [14] NDNP Technical Guidelines and Specifications. Available at: <http://www.loc.gov/ndnp/guidelines/>
- [15] ISO. Information and documentation – WARC file format (ISO 28100:2009), 2009.
- [16] OASIS Emergency Interoperability. OASIS. DOI=
<http://www.oasis-emergency.org>; Data Documentation Initiative (DDI). Data Documentation Initiative Alliance. DOI= <http://www.ddialliance.org/>
- [17] See for example, Clay Shirkey's "NDIIPP-Library of Congress Archive and Ingest Handling Test Report" (2005)
http://www.digitalpreservation.gov/partners/.../ndiipp_aiht_final_report.pdf
- [18] Hub and Spoke Project (HandS). UIUC. DOI=
<http://dli.grainger.uiuc.edu/echodep/hands/index.html>; Toward Interoperable Preservation Repositories (TIPR). FCLA. DOI= <http://wiki.fcla.edu:8000/TIPR/>; BagIt. CDL. DOI= <https://confluence.ucop.edu/display/Curation/BagIt>

Blogs as Objects of Preservation: Advancing the Discussion on Significant Properties

Karen Stepanyan

Department of Computer Science,
University of Warwick, Coventry,
CV47AL, UK

K.Stepanyan@warwick.ac.uk

Yunhyong Kim

School of Humanities, University of
Glasgow, UK
Yunhyong.Kim@glasgow.ac.uk

Matthias Trier

Department of IT Management,
Copenhagen Business School,
Denmark

mt.itm@cbs.dk

George Gkotsis

Department of Computer Science,
University of Warwick, Coventry,
CV47AL, UK

G.Gkotsis@warwick.ac.uk

Alexandra I. Cristea

Department of Computer Science,
University of Warwick, Coventry,
CV47AL, UK

A.I.Cristea@warwick.ac.uk

Hendrik Kalb

Institute for Business Informatics,
Technische Universität Berlin,
Germany

Hendrik.Kalb@ tu-berlin.de

Mike Joy

Department of Computer Science,
University of Warwick, Coventry,
CV47AL, UK

M.S.Joy@warwick.ac.uk

Seamus Ross

Faculty of Information, University of
Toronto, Canada

Seamus.Ross@utoronto.ca

ABSTRACT

The quest for identifying ‘significant properties’ is a common challenge for the digital preservation community. While the methodological frameworks for selecting these properties provide a good foundation, a continued discussion is necessary for further clarifying and improving the available methods. This paper advances earlier work by building on the existing InSPECT framework and improving its capabilities of working with complex/compound objects like blogs. The modifications enable a more thorough analysis of object structures, accentuate the differences and similarities between the framework’s two streams of analysis (i.e. Object and Stakeholder analysis) and, subsequently, improve the final reformulation of the properties. To demonstrate the applicability of the modified framework, the paper presents a use case of a blog preservation initiative that is informed by stakeholder interviews and evaluation of structural and technological foundations of blogs. It concludes by discussing the limitations of the approach and suggesting directions for future research.

Categories and Subject Descriptors

H.3.7 Digital Libraries

General Terms

Design, Theory

Keywords

Blogs, Weblogs, Digital Preservation, Significant Properties

1. INTRODUCTION

With the increasing number of blog-like services that encourage the propagation of user-generated content, the notion of a blog is becoming increasingly blurred [1]. However, developing an understanding of a blog as an information object is invaluable, especially within the context of preservation initiatives that aim to capture the authenticity, integrity and usability of blogs.

The ephemeral nature of web resources encouraged the development of long-term accessibility and preservation actions such as the Internet Archive¹ or HTTP Archive². Web archiving initiatives, such as ArcOMEM³ or LiWA⁴, have been increasingly trying to create solutions for social media archival situations. However, current preservation initiatives do not make adaptive provisions for dynamic and interactive environments such as blogs and social networking media. Instead, they tend to focus on various levels of version control and neglect deeper interactive aspects coming from networks, events and trends. This paper positions the conducted study within the context of blog preservation by highlighting the limitations of the current practices and emphasizing the rationale for developing blog preservation solutions. It demonstrates the pressing need to identify the properties of blogs that need to be preserved prior to embarking on a task of preservation. The paper proceeds to highlight the limitations within existing research on identifying these properties and proposes improvements accordingly. The paper concludes by demonstrating the application of the modified approach on a use case and discussing the benefits and limitations of the proposed approach.

¹ <http://archive.org>

² <http://httparchive.org/>

³ <http://www.arcOMEM.eu>

⁴ <http://liwa-project.eu>

2. RELATED WORK

As other Web resources, blogs are not immune from decay or loss. Many blogs that described major historic events, which took place in the recent past, have already been lost [2]. Another example that justifies preservation initiatives is the account of disappearing personal diaries. Their loss is believed to have implications for our cultural memory [3]. The dynamic nature of blogging platforms suggests that existing solutions for preservation and archiving are not suitable for capturing blogs effectively. However, blog preservation is not a trivial task.

Hank and her colleagues [4, 5] stress a range of issues that may affect blog preservation practices. The primary challenges of blog preservation are bound to the diversity of form that blogs can take and the complexity they may exhibit. A brief review of the literature shows that the definitions of blogs vary widely. The Oxford English Dictionary definitions of the terms ‘blog’ and ‘blogging’ highlight the temporal nature and periodic activity on blogs. Focus on technical elements of blogs is evident in the works by Nardi and his colleagues [6, p. 43]. Other definitions, for instance by Pluemavarn and Panteli [7, p. 200], deviate from a standpoint that looks into the technical aspects of blogs and into the socio-cultural role of blogs. The capacity of blogs for generating social spaces for interaction and self-expression [8] is another characteristic. The social element of blogs entails the existence of networks and communities embedded into the content generated by bloggers and their readership.

Due to the complexity of the Blogosphere - as shaped by the variety of blog types, the changing nature of blog software and Web standards, and the dependency on third party platforms - it is likely that lossless preservation of blogs in their entirety is unrealistic and unsustainable. Blog preservation initiatives should, therefore, question what essential properties they must retain to avoid losing their potential value as information objects. It becomes eminent that gaining insight into the properties of blogs and their users is necessary for designing and implementing blog preservation systems.

The quality of the preserved blog archives is dependent on capturing the fundamental properties of blogs. The following question would then be: what methods should be used for identifying these properties?

2.1 What to preserve in blogs: significant properties

In the digital preservation community, one of the prevailing approaches for defining what to preserve is bound to the notion of significant properties⁵ [9]. It is argued [10] that significant properties can help define the object and specify what to preserve, before deciding how to preserve. It has been acknowledged [11], however, that defining the significant properties without ambiguity remains difficult. The main problem is the lack of a suitable methodology for identifying the significant properties. While there are tools and frameworks for defining and recording technical characteristics of an object, Low [12] argues that identifying significant properties in general still remains contended, primarily due to the methods employed for the task. Low (*ibid.*) outlines the list of projects that attempted to develop mechanisms for identifying significant properties. The outcomes of these projects led to a range of frameworks and methodological

tools, such as PLANETS⁶ Plato that focuses on stakeholder requirements [13], InSPECT that combines object and stakeholder analysis [14], a JISC⁷-funded initiative that continues the discussion [15], and a template of characteristics [16] developed by NARA⁸.

Yet, despite the seemingly large number of tools that exist for organising significant properties into a range of types, expressing them formally, and testing their fidelity when subjected to selected operations (such as migration and emulation), the approaches available for guiding the decision making processes in identifying the relevant types and properties remain too abstract, especially with respect to complex objects [17].

However, considering the range of available solutions, InSPECT framework [14] is considered to offer a more balanced approach to identifying significant properties [12]. The advantage of this approach is encapsulated in the parallel processes it offers for analysing both the object and the stakeholder requirements. The framework is claimed to support identification of the significant properties of information objects by progressing through a specified workflow.

The InSPECT framework stands out as one of the first initiatives to accentuate the role of object functions derived from an analysis of stakeholder requirements as a gateway to identifying significant properties of digital objects.

2.2 Limitations of the Base Framework

InSPECT [14] is built on the Function-Behaviour-Structure framework (FBS) [18] developed to assist the creation and redesign of artefacts by engineers and termed useful for identifying functions that have been defined by creators of digital objects. The workflow of InSPECT is composed of three streams: Object Analysis, Stakeholder Analysis, and Reformulation. Each of these streams is further divided into stages that are argued by the authors (*ibid.*) to constitute the process of deriving significant properties of a preservation object.

However, the InSPECT framework was originally developed in line with simple objects such as raster images, audio recordings, structured text and e-mail. The main limitation of the framework, as discussed by Sacchi and McDonough [19], is its reduced applicability for working with complex objects. They (*ibid.*, p. 572) argue that the framework lacks “the level of granularity needed to analyze digital artifacts that — as single complex entities — express complex content and manifest complex interactive behaviors”. Similar complexities exist in the context of blogs, making application of InSPECT in its current form challenging. Hence, we propose a set of adjustments into the framework to improve its capability of working with objects like blogs.

The Object and Stakeholder Analysis are considered to be the two parallel streams termed as Requirements Analysis. Each of the streams results in a set of functions that are cross-matched later as part of the Reformulation stage. To address the limitation of InSPECT, we first focus on the lack of detailed instructions for conducting Object Analysis. The framework suggests the possible

⁶ <http://www.planets-project.eu/>

⁷ www.jisc.ac.uk/

⁸ <http://www.archives.gov/>

⁵ <http://www.leeds.ac.uk/cedars/>

use of characterisation tools or technical specifications for the purpose of object structure analysis (Section 3.1 of [12]). These suggestions presuppose the existence of such a tool or specification. While such a tool or specification may be available for fairly simple self-contained digital objects, like electronic mail, raster images, digital audio recordings, presentational markup, the situation is less straightforward for complex digital objects, such as weblogs and/or other social network media. In addition to the lack of guidance in defining the object structure, the framework suggests identifying functions associated with object behavior as part of the object analysis. These functions are then proposed to be consolidated with those identified from the stakeholder analysis stream. Consideration of functions introduces an ambiguously defined stakeholder view as part of the object analysis. This ambiguity and a higher level of abstraction when working with functions leads us to propose modifications of the framework to enable its application in the context of blog preservation.

3. PROPOSED CHANGES TO PRESERVATION PERSPECTIVES

The modifications discussed in this paper, firstly, introduce an ontological perspective into the Object Analysis stream and, consequently, further clarify the degree of overlap between the two streams of analysis. Secondly, it proposes integrating results from two separate streams at the level of properties rather than functions. We elaborate the proposed changes further down in this paper. We justify the changes introduced into the Object Analysis stream and clarify the subsequent adjustments to the workflow of the framework in the remaining part of this section. We then demonstrate the application of the framework by presenting a use case on blogs and discuss our experience in employing this approach.

3.1 Benefits of Ontological Perspectives

The modifications introduced in the Object Analysis stream aim to address the limitation of InSPECT (i.e. base framework) in specifying appropriate procedures for performing the analysis of complex objects and identification of their properties. We propose adopting an ontological perspective, to eliminate the impediment of the framework for guiding the preservation of objects such as blogs. Unlike simpler objects of preservation, such as images or text documents, blogs are usually comprised of other objects or embedded elements and demand a more structured approach when analysing these to avoid overlooking important properties.

The use of ontological perspectives is common in data modelling and has recently been receiving attention in the area of digital preservation. For instance, Doerr and Tzitzikas [20] refer to a set of ontologies, such as DOLCE, OIO and CIDOC CRM, established and commonly used in (digital) libraries, archives and related research initiatives. They (*ibid.*) argue that the use of ontologies makes the process of understanding sensory impressions of information objects more objective. Indeed, an ontological perspective can enhance the process of object analysis by offering abstraction to the level of conceptual objects along with the formalism for describing the structures of the compound objects. In contrast to current digital preservation research, Doerr and Tzitzikas (*ibid.*) emphasise the possible range of information objects (and relevant features) encompassed within a single information carrier and argue for exploring the sensory impressions rather than the binary forms objects. However,

stakeholder views are not directly discussed in the work by Doerr and Tzitzikas (*ibid.*). We attempt to follow Doerr's suggestion and integrate it with InSPECT. This enables us to use an ontological perspective for exploring complex objects (i.e. identifying compound objects and relationships among them) in addition to conducting a stakeholder analysis. The two streams of analysis can then be consolidated to inform the preservation strategy.

3.2 Description of the Framework

This section outlines each stage of the workflow and describes the major streams of analysis in greater detail. We focus on the modified parts of the framework, referring the reader to documentation of InSPECT for further details.

The diagrammatic representation of the proposed framework is presented in Fig. 1. The workflow of the framework initiates with the selection of the object of preservation and proceeds, via two parallel streams, to guide the Object and Stakeholder Analysis. The Object Analysis aims to establish the technical composition and the structure of the preservation object. This stage starts with the analysis of object structure. It focuses on the essence of object of preservation and aims to identify both conceptual and logical components of this compound object (viewed as classes). The next stage focuses on identifying relationships between the identified components. The relationships that exist between object components are expected to be explored and documented at this stage. Once the components and the relationships between those are identified, the properties of the object can be elicited and documented. The properties of the objects of preservation have to capture the characteristics of the compound objects along with their technical specifications. The stream of Object Analysis is therefore expected to result in developing a set of documented compound objects and associated properties that are to be cross-matched and refined with the outcomes of the parallel stakeholder analysis stream.

The Stakeholder Analysis aims at identifying a set of functions that stakeholders may be interested in and, subsequently, derive the properties of the preservation object that would be necessary to capture for supporting the required functions. The analysis starts with the identification of stakeholder types. They can be discovered through the investigation of policies, legal documents or communities related to the object. This stage is followed by the contextualisation of the object, which highlights stakeholders' perceived differences or variations in the levels of object's granularity. The third stage aims to determine the behaviour, which can be accomplished by examining the actions taking place in the real world. Having identified the actual behaviour, the anticipated behaviour is recorded through a set of functions. The last stage of the stakeholder analysis enables eliciting the properties of the object that are essential for satisfying the stakeholder requirements. The following stage aims at assessing and cross matching the properties identified from the two parallel streams of Object and Stakeholder Analysis.

The process of Cross-Matching and Refinement enables the consolidation of the identified properties and their refinement into an extended list of properties. The consolidation of the two independent streams is proposed to be conducted at the level of properties (rather than functions) and aims at integration of identified properties. The refinement of the integrated list of

properties leads to the proposal of properties to be considered for preservation. As significance is (repeatedly) associated with stakeholder views [21] the outcomes of the stakeholder analysis should remain in constant focus. The refinement of the integrated list should prioritise the inclusion of properties identified from the Stakeholder Analysis stream.

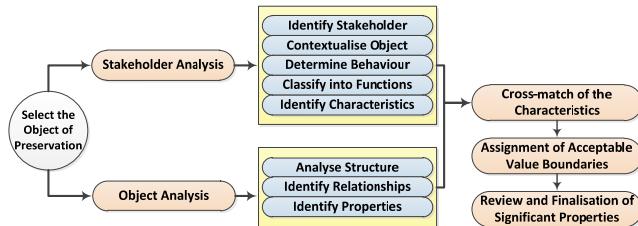


Fig. 1: Modified version of the base framework.

The Review and Finalisation stage includes the reflection on the previous steps and consideration whether any revisions are necessary. At this stage, identified properties can be recorded and the boundaries of their values can be assigned. The properties can then be used to define the objects of preservation and to progress with the design and development of the preservation initiative (for instance, for developing the databases necessary for storing data).

4. Use Case: Blog Preservation

This section integrates and consolidates some of the work carried forward as part of a blog preservation initiative [22, 23]. It describes the process of Object Analysis conducted to explore the object of preservation and (in the subsequent section) Stakeholder Analysis from the interviews exploring anticipated functionality of a blog repository.

4.1 Object Analysis

Blogs exhibit a considerable diversity in their layout, writing style or organisation. The analysis of this complex object, therefore, can be conducted from various perspectives and at different levels. Object analysis can employ an approach, widely accepted within the preservation community, that describes an information object as a conceptual (e.g., as it is recognised and perceived by a person), logical (e.g., as it is understood and processed by software), and as a physical object (e.g., as a bit stream encoded on some physical medium) [24]. In this section we present our approach adopted for the case of blogs and discuss this experience in a broader context. Identification of generic concepts of an object, their compound structures, hierarchy and relationships (without necessarily reflecting the operations expected to be performed) is common in ontology and data modelling. It can be used for the identification of generic concepts, subsequently leading towards the identification of object's properties [25]. A structured and iterative approach was adopted, to review and refine the analysis of the blog object. An alternative to this approach would involve consideration of an existing ontology. In this case, we conducted the following: [a] an inquiry into the database structure of open source blog systems; [b] an online user survey (900 respondents) to identify important aspects and types of blog data in the current usage behaviour; [c] suggestions derived from recent developments and prospects for analysing networks and dynamics of blogs; [d] an inquiry into the technologies, formats and standards used within the Blogosphere;

[e] an inquiry into blog structure based on evaluation of blog feeds (2,695 in total); and [f] an inquiry into blog APIs.

As a result of the above mentioned inquiries, a coherent view on the concepts of the blog object was acquired, informing further development of a respective data model. It enabled understanding the structure of blogs and help identifying their components, relationships and properties. The rest of this section outlines the process of conducting object analysis. Given the space limitation, a complete account of the performed study is omitted from this paper. We briefly outline the conducted work, the details of which are available elsewhere [see 23].

4.1.1 Database Structure, User Views and Network Analysis.

The knowledge of the domain, user survey and inquiry into conceptual models of blogs and their networks enabled identifying the most prominent conceptual and logical objects. Blogs may contain Entries (identified as being either Posts or Pages) that may have Comments and are associated with an Author. Both Entries as well as Comments exhibit certain Content. These entries are analysed further and (where relevant) broken down into smaller compound objects. For instance Content, as one of the most complex elements is further described by simpler objects like Tags, Links, Text, Multimedia, etc.. For demonstration purposes, we use only most frequently occurring components that are: Entry (Post/Page), Comment, Content, Author and the Blog itself, omitting the details due to space constraints.

In addition to the identification of compound entities of the complex objects, it is necessary to study the relationships that exist across these entities. This is particularly relevant when working with blogs, which are known to become interaction platforms and weave into complex networks. The structural elements of blogs, as conceptual, logical or physical objects, can represent the nodes and attributes, or define the ties of various networks. An insight into the network structures within and across blogs can be important gaining insight into the conceptual and logical objects. Identification of properties that may be of interest to archivists can greatly benefit from an insight into the network aspects of blogs and their components.

For instance, identifying different ways of citations within blogs can provide insight into the inter-related structure of objects, such as entries, comments or authors. However, while links added across blog posts may be technically similar to those added via link-back mechanisms, the ties formed by these two different types of links may be approached or treated differently. Our experience with this use case infers that the analysis of a blog in relation to others provides information about the properties of blogs and becomes useful as part of the Object Analysis stream. Furthermore, the theoretical and technological advances of analysing blogs and their networks should also be considered for gaining insight into the blogs and the phenomenon of blogging in general.

4.1.2 Technologies, Formats, RSS Structure and APIs.

While identification of compound elements and understanding of their relationships is an important step, it constitutes a high level view. To continue the analysis of the object and identify potential properties for preservation, a lower level view on the conceptual and logical objects is necessary. An inquiry into technical aspects

of blogs provides information about the lower level properties of the affiliated objects. To demonstrate this in the context of this use case, we highlight some examples of eliciting the properties of the blogs components.

To discuss an example of lower level properties we could consider the textual content. Textual content can be represented as a set of characters, along with its stylistic representation (e.g. font, colour, size), encoding, language, and bit stream expressed on the selected medium. The lower level description primarily deals with files, and can inform their storage and retrieval. Therefore, analysing the HTML code of blogs can reveal details about the technological backbone of blogs (formats, technologies, standards), which remains invisible to most blog users. Empirical studies exploring the physical objects can be particularly helpful in identifying potential properties. We briefly outline an example of a study to demonstrate the relevance of this approach.

Within the context of this paper, an evaluation of 209,830 blog pages has been performed [26]. The HTML-based representation of these resources was parsed and searched for specific mark-up used to define character sets, languages, metadata, multimedia formats, third-party services and libraries. The quantitative analysis of certain properties exhibited by the specific objects allowed us to describe common properties exhibited in blogs within the Blogosphere.

The evaluation was particularly useful in identifying properties of various compound objects (e.g. Content, which was further broken down into smaller logical objects and respective characteristics of associated physical ones). Geographical location (GPS positioning), as a contextual characteristic associated to Blog Entries or Content, was another direct outcome that emerged from the above evaluation. For instance, properties identified for the object Entry, and used in for demonstration purposes in this use case, include: [a] Title of the entry; [b] Subtitle of the entry; [c] Entry URI; [d] Date added; [e] Date modified; [f] Published geographic positioning data; [g] Information about access restrictions of the post; [h] Has a comment; [i] Last comment date; and [j] Number of comments. A more detailed description of the conducted analysis, as well as the complete list of objects and properties is made available elsewhere [23] due to space constraints.

4.2 Stakeholder Analysis

The objective of the Stakeholder Analysis is to identify a set of functions that stakeholders may be interested in and, subsequently, derive the properties of the preservation object that would be necessary to capture for supporting the required functions. The initial task was to identify or acknowledge the stakeholders that may interact with an instance of the object of preservation or their collection as part of a repository. Stakeholder interviews for identifying their requirements are an essential part of Stakeholder Analysis. Their methodological foundations as well as the complete list of functional requirements is available elsewhere [22]. A brief outline of the process directly affecting this use case is presented below.

4.2.1 Identification of Stakeholders.

Within the context of blog preservation we acknowledge three groups of stakeholders: Content Providers, Content Retrievers and Repository Administrators. Within each of these groups we identified individual stakeholders: [a] Individual Blog Authors;

[b] Organizations within the Content Providers group; [c] Individual Blog Readers; [d] Libraries, Businesses; [e] Researchers within the Content Retrievers group; and finally, [f] Blog Hosts/Providers and [g] Organizations (as libraries and businesses) within the Repository Administration group. This extensive list of stakeholders can be justified by the multitude of ways (including some unknown ways) of using preserved objects by present and future users [27]. Hence, rather than selecting a single definitive solution, it remains important to identify a range of essential as well as potential requirements to maximize the future usability of a blog repository. A user requirement analysis was performed for every stakeholder type. It focused on analysing stakeholder interaction with blogs via digital repository software.

4.2.2 Applied Method of Requirement Analysis.

There is a range of methods for conducting effective user requirement analysis [28]. In the context of this study we conducted an exploratory, qualitative study by means of semi-structured interviews. A set of stakeholders, from each of the groups, was approached to be interviewed. The structure of the interviews was designed to enable consolidation of the results across the stakeholders and stakeholder groups. General methodological and ethical guidelines for conducting qualitative inquiry of this kind were followed.

A total of 26 interviews were conducted. Candidate interviewees were identified and approached individually. The sample of interviewees was selected in a way that each of the defined stakeholder groups was represented by at least one interviewee. The distribution of interviewees for each of the stakeholder groups was: 10 for Content Providers; 12 for Content Retrievers; and 4 for Repository Administrators. The requirements were then analysed and a set of user requirements was identified.

4.2.3 Identified Requirements and Properties.

The analysis followed a three-step approach. Initially, each interview was analysed regarding the indication of requirements in the two main categories functional and non-functional. The non-functional requirements were classified into: user interface, interoperability, performance, operational, security and legal requirements. Subsequently, the requirements were analysed for recurrent patterns, aggregated and further clarified. The final list of identified requirements included a list of 115. Further details discussing the methods and the complete list of elicited requirements is available elsewhere [22]. The requirements that depend on existence of certain data elements were then shortlisted as shown in Table 1.

Table 1: A sample list of requirement functions identified from stakeholder interviews. (*FR: Functional Requirement, EI: Interface Requirements, UI: User Requirements, RA: Reliability and Availability Requirement)

Req.	Description	Req. Type*
R12	Unique URI with metadata for referencing/citing	FR/UI
R17	Distinguish institutional/corporate blogs from personal blogs	FR
R18	Display blog comments from several sources	FR
R19	Display and export links between/across blog content	EI/UI
R20	Prominent presentation of citations	FR/UI
R22	Historical/Chronological view on a blog	UI

Identifying data elements that are necessary for the implementation of the requirements leads to properties of the preservation object that can be attributed as important. Hence, the requirement analysis, in this case, proceeded in identifying data elements and conceptual entities they are associated with. The identified data elements are presented in Table 2. The properties elicited from the Stakeholder Analysis were then cross-matched with those resulting from Object Analysis stream and further refined into a consolidated list of properties.

Table 2: Properties elicited from stakeholder requirements.

Req.	Objects	Identified Properties
R12, R20	Entry	Digital Object Identifier(DOI)/Unique Identifier(UI)
R17	Blog	Blog type
R18	Comment	Comment type, source URI, service name
R19	Content	URI, URI type (e.g. external/internal)
R22	Blog, Entry, Comment	Creation/Capture/Update dates and time, time zone, date/time format.

4.3 Cross-Matching and Refining Properties.

The next step towards consolidating the list of properties includes the process of cross-matching, integration and refinement. The properties, identified from the two streams of Object and Stakeholder analysis are being compared and integrated into a single set of properties. It requires cross-matching and integration of properties that were missing from either of the list and eliminating same properties that were listed with different names.

We bring an example of cross-matching by referring to the property of DOI/UI9 for an entry, which has been identified from Stakeholder Analysis, but did not surface in Object Analysis. Unlike URIs that also constitute a unique identifier, an alternative approach similar to DOI was identified as necessary from the Stakeholder Analysis. Offering a consistent approach to referencing that is detached from the original resource differentiates between these identifiers. Hence, DOI/UI constitutes a property that is necessary for developing a system that meets stakeholder requirements. As a result, the property is added to the integrated list. This example demonstrates that Stakeholder Analysis allowed complementing the Object Analysis stream, which remained confined to intrinsic attributes of an entry such as URI.

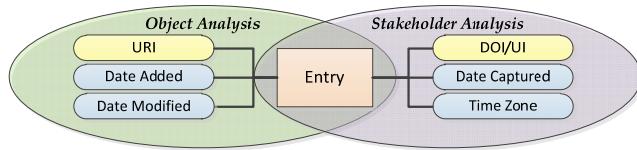


Fig. 2: An example of cross-matching and integration of properties, which were identified from the two parallel streams of Object and Stakeholder Analysis.

The requirement for providing a historical/chronological view of the entries, demonstrates another example where in addition to having the date and time of publication/editing, information about the time zone and date of capture is shown to be important. This can be elicited from the requirement R22 as shown in Table 2.

⁹ <http://www.doi.org/>

While dates have already been identified from the object analysis, their alignment within the repository that takes into account the time zone differences has been identified as important from the stakeholder analysis. The examples of cross-matching and integration are illustrated in Fig. 2.

4.3.1 Review and Finalisation of Properties

The final stage of the framework suggests to review the information collected at the previous stages and to decide whether additional analysis is necessary. The process of the review can be considerably improved if acceptable value boundaries are assigned to the identified properties. For instance, in line with the previous example, acceptable values and recognized standards can be considered for capturing the Time Zone and Date. Reflecting on acceptable boundaries can attest to the need for breaking down compound properties or reviewing the properties before their finalisation. The less abstract the identified properties are, the easier it would be to progress to the implementation of the preservation initiative. Returning to the Stakeholder Analysis and shortlisted requirements can reaffirm the identified properties or lead to further extension.

5. DISCUSSION

The use case (Section 4) represents an example of applying a methodological framework and informing a blog preservation initiative. It enables us to advance the discussion on identifying significant properties of complex objects such as blogs. Reflecting on our experience of the process of identifying and consolidating the object properties we report the benefits and disadvantages of employing this framework and suggest directions for further research.

The integration of the ontological perspective into the Object Analysis stream of the framework has indeed enabled a thorough analysis of the compound object under study. The results of object analysis produced a fine grained representation of the compound blog object. Integration of the ontological perspective into the InSPECT framework provided the lacking methodological guidance for working with complex objects. Furthermore, the modification of the framework that enabled cross-matching Object and Stakeholder Analysis streams at a lower level of properties has also been demonstrated beneficial. It clarified the process of comparison due to the use of specific properties rather than more abstract (higher level) functions.

However, the modified approach still lacks unambiguous methodological guidance for defining significance associated with each of the identified property. Supporting the identification of properties that are not significant will also be a useful addition to the framework. Potential directions for future work may involve developing tools for guiding stakeholder analysis and defining the levels of significance associated with properties. Exploring the possibilities of discussing the concept of significance as a relative spectrum should also be followed as part of the future research.

6. CONCLUSIONS

This paper advances the discussion on the topic of significant properties that engages the preservation community. It positioned the conducted inquiry within the context of blog preservation. Highlighting the limitations of current approaches in preserving blogs, this paper defined the rationale for understanding and defining blogs as objects of preservation.

Building on the body of work that provides methodological foundations for identifying significant properties, this paper adapted the recently developed InSPECT framework [12] for enabling its use with complex objects. It proposed to employ an ontological perspective on analysing compound objects enabling systematic analysis and de-composition of blogs into components and understanding the relations between them. This approach was demonstrated to be beneficial, leading towards identification of compound entities and properties of blogs. The modifications provided further clarification into the streams of Object and Stakeholder Analysis. Instead of cross-matching the functions, the framework proposes to consolidate the results at a lower and more tangible level of properties. While the use case demonstrated the applicability of the modified framework on the complex blog objects, it also highlighted a number of limitations. More specifically, further clarification is necessary for identifying properties that should not be considered for preservation. The development of methodological tools for defining and measuring significance is particularly important. Future work can also extend the discussion on automating the process of identifying these properties. The reuse and integration of existing ontologies is another direction that requires further examination. Nevertheless, the results emerging from the study summarised in this paper support the argument that the proposed modifications enhance the base framework by enabling its use with complex objects, and provide insight for advancing the discussion on developing solutions for identifying significant properties of preservation objects.

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

7. REFERENCES

- [1] Garden, M. Defining blog: A fool's errand or a necessary undertaking. *Journalism*(20 September 2011 2011), 1-17.
- [2] Chen, X. Blog Archiving Issues: A Look at Blogs on Major Events and Popular Blogs. *Internet Reference Services Quarterly*, 15, 1 2010), 21-33.
- [3] O'Sullivan, C. Diaries, on-line diaries, and the future loss to archives; or, blogs and the blogging bloggers who blog them. *American Archivist*, 68, 1 2005), 53-73.
- [4] Sheble, L., Choemprayong, S. and Hank, C. *Surveying bloggers' perspectives on digital preservation: Methodological issues*. City, 2007.
- [5] Hank, C. *Blogger perspectives on digital preservation: Attributes, behaviors, and preferences*. City, 2009.
- [6] Nardi, B. A., Schiano, D. J., Gumbrecht, M. and Swartz, L. Why we blog. *Communications of the ACM*, 47, 12 2004), 41-46.
- [7] Pluemavarn, P. and Panteli, N. *Building social identity through blogging*. Palgrave Macmillan, City, 2008.
- [8] Lomborg, S. Navigating the blogosphere: Towards a genre-based typology of weblogs. *First Monday*, 14, 5 2009).
- [9] Hedstrom, M. and Lee, C. A. *Significant properties of digital objects: definitions, applications, implications*. Luxembourg: Office for Official Publications of the European Communities, City, 2002.
- [10] Deken, J. M. Preserving Digital Libraries. *Science & Technology Libraries*, 25, 1-2 (2004/11/29 2004), 227-241.
- [11] Knight, G. and Pennock, M. Data without meaning: Establishing the significant properties of digital research. *International Journal of Digital Curation*, 4, 1 2009), 159-174.
- [12] Tyan Low, J. *A literature review: What exactly should we preserve? How scholars address this question and where is the gap*. University of Pittsburgh, Pennsylvania, USA., City, 2011.
- [13] Becker, C., Kulovits, H., Rauber, A. and Hofman, H. *Plato: a service oriented decision support system for preservation planning*. ACM, City, 2008.
- [14] Knight, G. *InSPECT framework report*. 2009.
- [15] Hockx-Yu, H. and Knight, G. What to preserve?: significant properties of digital objects. *International Journal of Digital Curation*, 3, 1 2008), 141-153.
- [16] NARA. *Significant Properties*. NARA, 2009.
- [17] Farquhar, A. and Hockx-Yu, H. Planets: Integrated services for digital preservation. *International Journal of Digital Curation*, 21, 2 2007), 88-99.
- [18] Gero, J. S. Design prototypes: a knowledge representation schema for design. *AI magazine*, 11, 4 1990), 26.
- [19] Sacchi, S. and McDonough, J. P. *Significant properties of complex digital artifacts: open issues from a video game case study*. ACM, City, 2012.
- [20] Doerr, M. and Tzitzikas, Y. Information Carriers and Identification of Information Objects: An Ontological Approach. *Arxiv preprint arXiv:1201.03852012*.
- [21] Dappert, A. and Farquhar, A. Significance is in the eye of the stakeholder. *Research and Advanced Technology for Digital Libraries*2009), 297-308.
- [22] Kalb, H., Kasioumis, N., García Llopis, J., Postaci, S. and Arango-Docio, S. *BlogForever: D4.1 User Requirements and Platform Specifications Report*. Technische Universität Berlin, 2011.
- [23] Stepanyan, K., Joy, M., Cristea, A., Kim, Y., Pinsent, E. and Kopidaki, S. *D2.2 Report: BlogForever Data Model*. 2011.
- [24] Thibodeau, K. *Overview of technological approaches to digital preservation and challenges in coming years*. Council on Library and Information Resources, City, 2002.
- [25] Dillon, T., Chang, E., Hadzic, M. and Wongthongtham, P. *Differentiating conceptual modelling from data modelling, knowledge modelling and ontology modelling and a notation for ontology modelling*. Australian Computer Society, Inc., City, 2008.
- [26] Banos, V., Stepanyan, K., Joy, M., Cristea, A. I. and Manolopoulos, Y. *Technological foundations of the current Blogosphere*. City, 2012.
- [27] Yeo, G. 'Nothing is the same as something else': significant properties and notions of identity and originality. *Archival Science*, 10, 2 2010), 85-116.
- [28] Hull, E., Jackson, K. and Dick, J. *Requirements engineering*. Springer-Verlag New York Inc, 2010.

Challenges in Accessing Information in Digitized 19th-Century Czech Texts

Karel Kučera

Charles University in Prague, Czech Republic
nám. Jana Palacha 2
11638 Praha 1
00420 224241490
karel.kucera@ff.cuni.cz

Martin Stluka

Charles University in Prague, Czech Republic
nám. Jana Palacha 2
11638 Praha 1
00420 224241490
martin.stluka@ff.cuni.cz

ABSTRACT

This short paper describes problems arising in optical character recognition of and information retrieval from historical texts in languages with rich morphology, rather discontinuous lexical development and a long history of spelling reforms. In a work-in-progress manner, the problems and proposed linguistic solutions are shown on the example of the current project focused on improving the access to digitized Czech prints from the 19th century and the first half of the 20th century.

Categories and Subject Descriptors

H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – *linguistic processing, dictionaries*.

General Terms

Languages

Keywords

Information Retrieval, Known-Item Retrieval, Historical Text, Lemma, Hyperlemma

1. INTRODUCTION

As has been recently pointed out, in spite of undeniable progress over the last few years, the state-of-the-art software for optical character recognition still does not provide satisfactory results in transformation of historical books, magazines and newspapers into searchable and editable text. [1] Low quality of old prints, use of historical typefaces (such as the Gothic script in its numerous regional variants), special characters and ligatures, ageing of paper and page curl are usually mentioned among the major technical OCR difficulties being worked upon. However, the whole problem also has a linguistic aspect, since the results of OCR can be substantially improved by linguistic information, as has been proved in OCR of modern texts in tens of languages

where extensive language-specific lists of paradigmatic word forms have been used to optimize the OCR ‘best guesses’ by comparing the resulting interpretations of character strings to existing word forms.

Long overshadowed both by the abovementioned technical issues and the more urgent demand to achieve high dependability of OCR results in modern texts, the problems of using historical lexica in noisy old text data has been fundamentally addressed only lately [2]. At the same time, there has been designed a plausible way of building period-specific lexica from manually corrected ground-truth texts and/or from historical dictionaries (if available), [3] but so far few lexica have been compiled and tested in practice. One notable exception was the series of tests performed under the European IMPACT program, which included historical lexica for nine languages and showed that “deployment of historical lexica improves the state-of-the-art of both OCR and IR”. [4]

Generally speaking, the deployment of historical lexica for OCR and IR purposes should help to solve the language-related noise coming from

- archaic words (such as *eftsoon* ‘again; at once’ or *thine*, to give English examples) and word formations (*disobedency, labourous* etc.)
- archaic inflectional forms (e.g. *maketh, makest, bespeak*) and
- archaic spellings like *oeconomic, aeternal, to-morrow, applyed, fruitfull, hydraulick* etc.

To compile a historical lexicon may represent different degrees of challenge, depending on how numerous and complicated the differences from the present language are in the above three areas, as well as on some other factors such as the availability of dictionaries and grammars from the particular period or accessibility of computer processable editions of historical texts. Moreover, the challenge is different in different types languages: the compilation of a lexicon may be relatively trivial in predominantly isolating languages like English, where inflected words follow a very limited number of paradigms with a very limited number of forms in each of them, as compared to highly inflectional languages, with up to several tens of forms in each of

tens or even hundreds of paradigms diversified by grammatical categories, sound changes, variations and fluctuations.

In the following, we elaborate on the specific problems connected with the historical lexicon building in Czech, as they are approached in the project *Tools for Accessibility of Printed Texts from the 19th Century and the First Half of the 20th Century*. [5]

2. THE CASE OF CZECH

2.1 General Background

The texts from the 19th century and the first half of the 20th century which are in the focus of the aforesaid Czech project, are not too far removed from the present texts, and given the availability of several 19th- and 20th-century Czech dictionaries and grammars, it may seem to be a relatively unsophisticated task to compile a historical lexicon for OCR and IR purposes. At a closer look, however, the task is not quite as trivial, mainly due to historical reasons. At the beginning of the 19th century, German and Latin were the high-status languages in the Czech lands, while Czech was struggling for full-fledged existence, being practically unused in technical and scientific writing, ‘high’ poetry or prose. However, only 50 years later, following a vocabulary explosion, intensive de-Germanization and wide-ranging refinement resulting from the National Revival movement, the situation was completely different. Generally, this line of development continued, if in a less intensive way, in the second half of the 19th century, but while the Czech vocabulary kept growing in a number of branches of technology and science, more German loan words and many of the unsuccessful neologisms coined in the earlier period were being abandoned. Considering the modern Czech language of the 1st half of the 20th century, with its fully developed terminology and variety of language styles, one can conclude that at least three different lexica should be created to accommodate the OCR and IR needs, each covering a period of about 50 years.

Nevertheless, one more important factor needs to be taken into consideration in the Czech case, namely the three deep-cutting reforms of orthography implemented in 1809, 1843 and 1849, which changed the use of several high-frequency letters and, consequently, the spelling of tens of thousands of word forms. The following four spellings of the same example sentence (meaning ‘All this happened not by her fault, but someone else’s’) stand as telling samples of how pronounced the changes were:

- until 1809: *To wſſe ſe ſtalo ne gegj, ale cyzý winau.*
- until 1843: *To wſe ſe ſtalo ne gegj, ale cizj winau.*
- until 1849: *To wſe ſe ſtalo ne její, ale cizí winau.*
- after 1849: *To vše ſe ſtalo ne její, ale cizí vinou.*

As a consequence, four lexica, each of them reflecting different spellings and rather different vocabularies, are being worked on to cover the 150-year period. In fact, four more lexica will be compiled, each of them including both the pre-reform and post-reform spelling variants. These lexica will be used in OCR and IR with the prints from the short transitory periods when the orthographic reforms were only being introduced and the older and newer spellings were used in the same texts.

2.2 Building the Lexica

The compilation of each of the four Czech historical lexica is based on the combined use of lists of headwords obtained from 19th- and 20th-century dictionaries and/or lists of word forms extracted from available OCRed or manually transliterated historical texts. After a proofreading, the lists are processed in the following four steps:

- Each word form on the list is assigned a modern lemma, i.e. a citation/dictionary form written in modern spelling. Applying this approach, the English forms *make, makes, made, making* would be all assigned the lemma *make*; the modern lemma for historical spellings as *oeconomic, aeternal, to-morrow, applyed, fruitfull, hydraulick* would be *economic, eternal, tomorrow, apply, fruitful, hydraulic* etc. The unrecognized forms in all the lexica are reviewed and either discarded as noise or accepted, corrected (in the case of OCR misreadings) and manually lemmatized. The procedure for the words and word forms printed in one of the pre-1849 spellings is different in that they are first converted into modern spelling and only then (automatically or manually) assigned a lemma.
- The lemmata are then distributed into groups according to their paradigmatic characteristics, i.e. according to the way they inflect. Special attention is given to integrating all old forms (in English, for example, *maketh, makest*) into the paradigms.
- Using a paradigm-specific utility for each of the groups, the lemmata are expanded into full paradigms, many of which in the case of Czech include up to several tens of forms. The modern lemma accompanies each generated form, so that the resulting lines of the lexicon have the format “form;lemma”, i.e. for example *wilou;vila*.
- Finally, the full paradigms based on the transcribed pre-1849 spelling forms (cf. step one above) are converted back to the spelling identical with the one originally used. Depending on the original spelling, the line quoted as an example in the previous paragraph would then be changed in one of the following: *wjlau;vila* (pre-1843 spelling), *wilau;vila* (pre-1849 spelling) or *vilou;vila* (post-1849 spelling).

Ideally, the resulting initial versions of the lexica at this point include complete paradigms of all the words found in the texts and/or dictionaries used for their compilation. However, the lexica are paradoxically far from being ideal, especially from the IR viewpoint.

2.3 Reductions and Additions

Experience with the lexica compiled in the above-described way showed that some rare or unused items (mostly archaisms and neologisms) tend to penetrate into the them as a result of the fact that such words had their own entries in Czech 19th-century dictionaries. This, again, had its historical reasons: especially in the first half of the century, the author of a dictionary might wish not just to reflect the real usage, but also to show that the richness of the Czech vocabulary was comparable to that of German, which may have not been quite true then. As a result, the dictionary in fact partly demonstrated the potential of Czech by including new coinages and centuries-old words, not just the contemporaneous usage.

Experience also showed that the lexica are overgenerated, especially in that they include all the low-frequency forms of low-frequency words. Out of context, such comprehensiveness may be desirable, but in practice it proved counterproductive. In Czech, this is primarily the case of transgressive forms of low-frequency verbs, which may have never been used in Czech texts but are often homonymous with forms of other words, many of them high-frequency ones, such as for example *podle* (transgressive of the rare verb *podlít* ‘stay for a short time’) and *podle* (high-frequency preposition meaning ‘according to’ or ‘by’). As such, they are potential sources of noise in IR.

On the other hand, in the course of time, thousands of words and forms will have to be added to the initial versions of lexica which, with over 500,000 word forms in each of the four of them, are still somewhat limited as a natural result of the fact that a rather limited number of computer-processable texts and dictionaries were available for their compilation. New items will be added to the lexica from a growing number of texts in the following four years of the project. The general expectation is that most additions will come from technical texts and poetry, but there will no doubt be one more, rather specific group coming from the prose, press and drama that partly reflected the colloquial stratum of the Czech vocabulary of the 19th century. Characterized by hundreds of German loan words, this largely unresearched part of the Czech word-stock was mostly ignored in the 19th-century dictionaries owing to the anti-German linguistic attitudes prevailing during the Czech National Revival and the following decades.

The difficulties presented by the lexica including rare or unused words and forms on the one hand, and missing colloquial words and forms on the other, are different in OCR and IR. In OCR, problems arise if the missing words or the rare/unused words happen to be formally similar to (but not identical with) some common forms, because the similarity may cause OCR misinterpretations. Formal identity (i.e. homonymy) of two or more forms is irrelevant because what matters in OCR is the mere existence of the form, not its meaning(s) or grammatical characteristic(s).

In IR, on the other hand, homonymy is the main source of difficulties as it may cause a considerable increase in the amount of noise in the results of end-users’ queries. Formal similarity (not identity) of word forms itself does not present any direct problems for IR, but influences its results indirectly, through the abovementioned OCR misinterpretations.

To reduce these problems, a record will be kept of occurrences of words (lemmata) and their forms in the processed texts, with metadata including the ID of the text, page number and position of the word form on the page as well as information about the text including the year of its publication, text type (belles-lettres, press, science and technology) and domain (natural sciences, medicine, mathematics etc.). The reviewed record will be periodically used to add words and word forms to the existing lexica. Eventually, towards the end of the project it should also also be used for a realistic reduction of the initial lexica to words and forms attested in authentic texts. At the same time, the extensive record, estimated to include more than 5,000,000 word forms by the end of the project, should help to differentiate between generally used words and special vocabularies, as well as between words and forms used during the entire 150-year period and those with a limited life span.

3. LINGUISTIC INFORMATION AND IR

As shown above, in the Czech project the added linguistic information in the lexica consists in assigning a lemma to each word form. As a form representing the entire set of paradigmatic forms of a particular word, the lemma makes it possible to efficiently retrieve all the occurrences of all the forms of the searched word at once – a capacity especially appreciated by end-users performing searches in languages in which words may have numerous forms.

Assigning the correct lemma to all the word forms in the text can also help to remove many of the problems caused by homonymy: in this way, for example, the homonymy in the English *left* (‘opposite of right’ or past tense of the verb *leave*) can be eliminated. However, to assign the correct lemmata to homonymic words or word forms requires disambiguation, which in the case of historical texts can practically only be manual as, to our knowledge, there exist no acceptably functional historical disambiguation programs for old Czech or other old languages. Since manual disambiguation is far too inefficient in projects where the number of digitized and OCRed pages of old texts amounts to thousands a day, homonymy remains an interfering problem in IR. In the Czech case, for the time being, the homonymic forms are standardly assigned as many lemmata as many paradigms they are part of.

Nonetheless, if the strict linguistic definition of the lemma is stretched a little, the concept can accommodate more end-users’ needs than just the clustering of all the forms of a word. Dubbed as “hyperlemma”, the extended concept is being implemented in the ongoing lemmatization of the diachronic part of the Czech National Corpus, [6] representing not only the paradigmatic forms of words, but also their phonological and spelling variants used during the seven centuries of Czech texts. Thus, in a hyperlemma query, the user is free to use the modern phonological/spelling form of the lemma (e.g. *angažmá*, *téma*) to retrieve all the instances of its modern and historical forms and variants (in this case *engagement*, *engagementu*, *engagementem...*, *thema*, *thematu*, *thematem...*). The employment of the concept will arguably be even more important in the discussed Czech project than it is in the corpus, because unlike the corpus, the typical users of which are linguists, the body of texts which is in the focus of the project is expected to be used typically by historians and other scientists as well as by journalists and the general public, that is by people without a deeper knowledge of the historical changes in the Czech language.

In view of further problems they may experience when searching for a particular known item in the texts from the 19th and the first half of the 20th century, the following four general situations (and solutions) were considered:

- The word the user is searching for exists in just one phonological and spelling form used now as well as in the 19th century, and none of its paradigmatic, phonological or spelling forms overlaps with any form of any other word. The retrieved forms will be exactly those (s)he is looking for. This is the ideal (and, fortunately, also majority) case presenting no problems.
- The word the user is searching for exists in two or more modern phonological and/or spelling variants with the same meaning and about the same frequency (e.g. *sekera/sekyra* ‘ax’, *vzdechnout/vzdychnout* ‘to sigh’, the suffix

-ismus/-izmus ‘-ism’), or in two or more historical phonological and/or spelling variants of the same meaning and about the same frequency (*čív/číva* ‘nerve’). There are hundreds of such cases in Czech; in English this is a relatively rare phenomenon (e.g. *ax/axe*) unless one considers the multitude of British and American spelling variants such as *humour/humor*, *theatre/theater*, *materialise/materialize* etc. To avoid the problems caused by the rather common situation that the user may not realize the parallel existence of the variants and consequently will miss part of the searched-for information, a record of these variants is being built and used by the search program. After one of such lemmata is keyed in (e.g. *ax*), the program will automatically retrieve all the forms of all the variants (i.e. *ax, axe* and *axes*), and the user will be informed about it.

- The word the user is searching for exists in one or more common modern phonological and/or spelling variants, with the same meaning and about the same frequency (e.g. *anděl* ‘angel’, *myslet* ‘to think’) and in one or more infrequent or presently unused (mostly historical) variants of the same meaning (*anjel*, *myslit*). Many users will not be aware or think of the existence of the latter variant(s), so again, to avoid the risk of missing part of the searched-for information, a record of these variants is used, if in a slightly different procedure. The planned solution is that once the commonly/frequently used lemma (e.g. *anděl*) is keyed in, the search program will retrieve all the forms of all the lemmata (*anděl, anděla, andělovi, andělem..., anjel, anjela, anjelovi, anjelem...*), and the user will be informed about it. On the other hand, if the user keys in the currently unused/infrequent lemma (*anjel*, in this case), the program will only retrieve the forms of this lemma (i.e. *anjel, anjela, anjelovi, anjelem...*). The reasoning behind the latter procedure is that the user is obviously not a complete laymen, knows the form and has a reason to search for it. In case the user wants to retrieve just the forms of the more frequent variant (*anděl*), (s)he can revert to the string-matching query.
- The word the user is searching for only exists in one modern/historical phonological and spelling variant (i.e. it has one lemma), but one or more of its forms are homonymic, i.e. overlap with forms of another lemma, as in the example of *left* (‘opposite of right’ or past tense of the verb *leave*) given above. Czech as a highly inflectional language has thousands of such homonymic word forms, with some of them being part of four or even five different paradigms, and, as has been stated above, at present there is no practicable way to significantly reduce the noise such forms cause in IR from historical texts. A record of homonymic forms is being compiled for the future use in a disambiguator of historical Czech texts but in the nearest future its use will be mostly limited to informing the user about the problem whenever (s)he is searching for a lemma including homonymic forms.

4. CONCLUSION

While homonymy will remain one of the main problems of IR from historical texts in Czech as well as in many other languages, the expectation is that the results of the Czech project will make known-item retrieval easier for the end user, especially by implementing the abovementioned concept of hyperlemma and by modifying the query based on lists including both contemporary and historical variants. As a result, still on the linguistic ground, the user will be able to find, with a single query, all instances of all attested present and historical forms and spelling/phonological variants of a word – a feature which is not common in similar text collections (with very few exceptions like *encyclopedia* and *encyclopaedia*, several searches must be performed to find different forms like *go, goes, goeth; economy, oeconomy; medieval, mediaeval; peaceful, peacefull* etc. in Google books, Hathi Trust Digital Library, Open Library, the University of Michigan Collection and others). [7]

Last but not least, the lexica and lists being compiled under the Czech project will serve as a basis for the development of a disambiguator for the texts from the 19th century and the first half of the 20th century.

5. REFERENCES

- [1] *IMPACT 2011 Project Periodic Report*, 5, http://www.impact-project.eu/uploads/media/IMPACT_Annual_report_2011_Publishable_summary_01.pdf.
- [2] Gotscharek, A., Neumann, A., Reffle, U., Ringlstetter, C., and Schulz, K. U. 2009. Enabling information retrieval on historical document collections: the role of matching procedures and special lexica. In *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data*, AND '09. ACM New York, NY, 69-76.
- [3] Depuydt, K. 2009. Historical Lexicon Building and How it Improves Access to Text. (Conference *OCR in Mass Digitisation: Challenges between Full Text, Imaging and Language*. The Hague). See presentation at https://www.impact-project.eu/uploads/media/Katrien_Depydt_Historical_Lexicon_Building.pdf.
- [4] *IMPACT 2011 Project Periodic Report*, 11, http://www.impact-project.eu/uploads/media/IMPACT_Annual_report_2011_Publishable_summary_01.pdf.
- [5] Part of the of the *Applied Research and Development of National and Cultural Identity Programme (NAKI)* funded by the Czech Ministry of Education. For details see <http://www.isvav.cz/programmeDetail.do?rowId=DF> and <http://kramerius-info.nkp.cz/projekt-naki>.
- [6] See www.korpus.cz.
- [7] Searches performed at <http://books.google.cz/books>, <http://www.hathitrust.org>, <http://archive.org/details/texts>, <http://quod.lib.umich.edu/g/genpub?page=simple>.

ESA USE CASES IN LONG TERM DATA PRESERVATION

Mirko Albani

Rosemarie Leone

Calogera Tona

ESA-ESRIN
Via G. Galilei
CP 64,00044 Frascati
Italy
name.surname@esa.int

ABSTRACT

Long Term Data Preservation (LTDP) aims at ensuring the intelligibility of digital information at any given time in the near or distant future. LTDP has to address changes that inevitably occur in hardware or software, in the organisational or legal environment, as well as in the designated community, i.e. the people that will use the preserved information. A preservation data manages communication from the past while communicating with the future. Information generated in the past is sent into the future by the current preservation data. European Space Agency (ESA) has a crucial and unique role in this mission, because it maintains in its archives long time series of Earth Observation (EO) data. In order to ensure to future generations data use and accessibility of this cultural heritage is needed to define a systematic approach, accompanied by different use cases.

Keywords

Long Term Data Preservation (LTDP), Data Curation, ESA, EO data, Preserve Data Set Content (PDSC).

1. INTRODUCTION

The main objective of the Long Term Data Preservation (LTDP) initiative is to guarantee the preservation of the data from all EO ESA and Third Parties ESA managed missions on the long term, also ensuring their accessibility and usability, as part of a joint and cooperative approach in Europe aimed at preserving the EO European data from member states' missions [1].

The concept of LTDP can be characterized as communication with the future. In the future new technology will be used that is more cost effective and more sophisticated than current technology. Communication with the future then corresponds to moving records onto new choices of technology. The preservation environment will need to incorporate new types of storage systems, new protocols for accessing data, new data encoding formats, and new standards for characterizing provenance, authenticity and integrity. The long term preservation of Earth Observation data is a major issue today as monitoring of global change processes has led to increasing demand for long-term time series of data spanning 20 years or more also in support to international initiatives such for example the United Nations Framework Convention on Climate Change (UNFCCC), the ESA Climate Change Initiative (CCI) and the Global Monitoring for Environment and Security program (GMES). The large amount of new Earth Observation missions upcoming in the next years will lead to a major increase of EO space data volumes and this fact, together with the increased demands from the user community, marks a challenge for Earth Observation satellite operators, Space Agencies and EO space data providers regarding coherent data

preservation and optimum availability and accessibility of the different data products. The preservation of EO space data can be also in the future as a responsibility of the Space Agencies or data owners as they constitute a humankind asset.

In 2006, the European Space Agency (ESA) initiated a coordination action to share among all the European (and Canadian) stakeholders a common approach to the long term preservation of Earth Observation space data. During 2007, the Agency started consultations with its Member States presenting an EO Long Term Data Preservation strategy targeting the preservation of all European (including Canada) EO space data for an unlimited time-span ensuring and facilitating their accessibility and usability through the implementation of a cooperative and harmonized collective approach (i.e. a European EO LTDP Framework) among the EO space data owners in order to coordinate and optimize European efforts in the LTDP field and to ultimately result in the preservation of the Completed European EO space data set for the benefit of all European countries and users and with a reduction of overall costs.

The Long Term Data Preservation Working Group with representatives from ASI, CNES, CSA, DLR and ESA was formed at the end of 2007 within the Ground Segment Coordination Body (GSCB) [1], with the goal to define and promote, with the involvement of all the European EO space data and archive owners, the LTDP Common Guidelines and also to increase awareness on LTDP. The LTDP guidelines were published at the end of 2009 and constitute a basic reference for the long term preservation of EO space data. Their application by European EO space data owners and archive holders is fundamental in order to preserve the European EO space data set and to create an European LTDP Framework. The application of the identified guidelines is not a requirement or a must for European EO space data owners and archive holders but is strongly recommended also following a step-wise approach starting with a partial adherence.

This paper is organized as follows: Section 2 presents state of the art, Section 3 shows LTDP architecture, Section 4 provides a use cases overview and Section 5 presents the conclusions and future developments

2. PRESERVATION OF EO SPACE DATA: STATE OF THE ART

The main milestones in LTDP development are:

- LTDP Framework
- LTDP Common Guidelines

- LTDP Preserve Data Set Content

2.1 LTDP FRAMEWORK

LTDP Framework, and others such as European LTDP Common Guidelines, was produced by the LTDP working group. The concepts contained in this document were presented to the EO data owners and archive holders community at the 1st LTDP workshop held at ESA/ESRIN in May 2008 [2].

Its main goal is to define a “European LTDP Framework” aimed at providing a practical way of carrying on LTDP activities at European level. The initial concepts and ideas contained in this document should help the establishment of a European LTDP Framework to coordinate and optimize European efforts in the LTDP field, that, in turn, would ultimately result in the preservation of the complete European data set with a coherent and homogeneous approach for the benefit of all European countries and users and with a reduction of overall costs.

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

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

The European LTDP Framework is open to all possible members and is to be intended as a collaborative framework consisting of distributed and heterogeneous components and entities cooperating in several areas to reach a harmonized preservation of the European EO space data set. The framework is based on the contribution of European EO space data owners through their ideas and possibly their infrastructure in accordance to the commonly agreed LTDP Guidelines and should follow a progressive implementation based on a stepwise approach (short, mid, long-term activities). A common approach in the field of Long Term Data Preservation should aim at the progressive application of the European LTDP Common Guidelines but also at cooperation of the archive owners in several areas for a progressive development and implementation of technology, methodology, standardization, operational solutions and data exploitation methodologies as key aspects for the set-up of the framework.

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

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

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

2.2 LTDP COMMON GUIDELINES

The European LTDP Guidelines are intended to cover the planning and implementation steps of the preservation workflow and have been defined on the basis of a high-level risk assessment performed by the LTDP Working Group on the Preserved Data Set Content [3] and its composing elements.

The LTDP guidelines and the underlying data preservation approach should be applied not only to future missions, where they can be easily and systematically included in the mission operations concept starting from the early phases with consequent cost savings and better achievable results, but also to the missions currently in operation or already disposed. In those last cases their application and the recovery of the full EO PDSC content could be trickier and not completely achievable and tailoring might be necessary. For current and not operational missions in any case, an incremental application approach should be pursued; the approach should consist in auditing the archives versus the LTDP Guidelines and PDSC document to be followed by the implementation of the highest priority and by the recovery of critical missing data/information.

In the field of Earth Observation, the data landscape is complex and there will naturally be different user communities with divergent needs for the long term reuse of the data. In case a more specific designated user community has to be addressed wrt, more specific preservation objectives and PDSC content should be defined and the LTDP Guidelines might need to be refined and augmented accordingly. In those cases it is recommended to follow the steps using the PDSC and the LTDP guidelines as starting point for the definition of a more specific approach to be properly documented in the form of “preservation approach and strategy” documents.

2.2.1 *Preservation analysis workflow*

Preservation of Earth Observation data should rely on a set of preservation actions properly planned and documented by data holders and archive owners, and applied to the data themselves and to all the associated information necessary to make those data understandable and usable by the identified user community. Data holders and archive owners should follow the “Preservation Analysis Workflow” procedure to define the proper preservation strategy and actions for their Earth Observation data collections. The result of the procedure application should consist of a set of documents describing the preservation strategy, implementation plan and activities for each individual mission dataset. Such document(s) should refer to the LTDP guidelines and clearly

define current compliance and future plans to improve adherence. The procedure consists of the following steps:

- Definition of preservation objective and designated user communities.
- Definition of Preserved Data Set Content (PDSC) for Earth Observation missions.
- Creation of PDSC Inventory for each own EO mission/instrument dataset.
- Risk assessment, preservation planning and actions, risk monitoring.

These steps are applicable to any digital data repository and are shortly described below for a generic Earth Observation case:

The preservation objective considered here for an Earth Observation data holder and archive owner consists in maintaining the own full data holdings accessible and usable today and in future, theoretically for an unlimited time, for its designated user communities. Long-term accessibility and usability of Earth Observation data requires that not only sensed data but also the associated knowledge (e.g. technical and scientific documentation, algorithms, data handling procedures, etc.) is properly preserved and maintained accessible. This implies the availability and archiving of metadata and data products at all levels specified by each owned mission or the capability to generate them on user request through proper processing. Data products need moreover to be provided with known quality to end-users together with the information necessary to understand and use them.

Different designated user communities are addressed through the preservation objective defined above. Earth Observation data users are today, as an example and among others, Scientists and Principal Investigators, researchers, commercial entities, value adders, and general public. These communities can be further differentiated on the basis of the respective application domain and area of interest (e.g. ocean, atmosphere) and generally have different skills, resources and knowledge. The data product levels and the information associated to the data necessary for their understandability and use is different for each of the above communities and even for individuals inside each community. Earth Observation data holders and archive owners generally serve today more than one user community and therefore need to be able to address the needs of all of them in terms of data and associated information availability and access. In addition, the preservation objective includes the utilization of the data products also in the future by user communities that might have completely different skills and knowledge base wrt the ones identified today but also different objectives for the use of the data. This means that the best approach for Earth Observation data holders and archive owners today would be to consider a “designated user community” generic and large enough so that the identified content to be preserved in the long term for that community will allow also other users, not considered at the time preservation was initiated, to make use of the data in the future. The generic designated user community is assumed to be able to understand English, to work with personal computers and basic programs provided with them, and to analyse and interpret the data products when available together with the full amount of additional information necessary to understand them without additional support from the archive.

In Earth Observation, the “Preserved Data Set Content” should be comprised as a minimum, in addition to the EO data, of all

information which permit the designated user community to successfully interact, understand and use the EO data as mandated by the preservation objective. The Earth Observation Preserved Data Set Content has been defined on the basis of the preservation objective and generic designated user community.

For past and current missions, the next stage to be implemented by data holders and archive owners is to tailor the PDSC for each EO mission/instrument, and to appraise each of the resulting elements comprised in the preserved data set content in terms of physical state, location and ownership. The result is the mission/instrument inventory document. For future missions, the definition of the PDSC shall be initiated during the mission definition and implementation phases and continuously maintained in the following phases.

Risk assessment in terms of capability of preservation and accessibility for each element of the inventory should be then performed and the most appropriate preservation actions identified and planned for implementation. The result of this activity should consist of one or more “preservation strategy and approach” documents. These documents could be drafted with different levels of detail and should generally contain Preservation Networks for each EO mission data collection consisting of all the PDSC Inventory elements, the elements on which they are dependent or necessary to understand and use them (e.g. the operating system underlying an EO data processor) and the associated preservation actions identified for each of them. Preservation networks should also identify the preservation state of each element of the PDSC inventory. Such document(s) should refer to the LTDP guidelines and clearly define current compliance and future plans to improve adherence. The identified preservation actions should be then implemented and the risks associated with inventory elements preservation properly and continuously monitored.

2.2.2 LTDP Guidelines Content

The guiding principles that should be applied to guarantee the preservation of EO space data in the long term ensuring also accessibility and usability are:

- Preserved data set content
- Archive operations and organization
- Archive security
- Data ingestion
- Archive maintenance
- Data access and interoperability
- Data exploitation and re-processing
- Data purge prevention

The LTDP guidelines constitute a basic reference for the long term preservation of EO data. Their application by European Earth Observation space data holders and archive owners is fundamental in order to preserve the European EO space data set and to create a European LTDP Common Framework. The application of the identified guidelines is not a requirement or a must for European EO data holders and archive owners but is strongly recommended along with following a step-wise approach starting with a partial adherence. The key guidelines should be

intended as a living practice and as such might evolve following specific research and development activities (e.g. outcome of cooperation in LTDP in Europe). Each key guideline could also have associated a set of technical procedures, methodologies or standards providing technical details on the recommended practical implementation of the guideline. Their selection has been made considering the results of cooperation activities in Europe with the goal to favour convergence in Europe on the LTDP approach and implementation.

Similarly to the key guidelines, these procedures or standards could be further evolved and improved with time or even developed or defined if missing. This can therefore also be intended as a starting point to support the establishment, and aid the implementation, of such detailed procedures or methodologies when missing, favouring active cooperation in Europe in the LTDP field. LTDP principles and key guidelines considered necessary to initiate this process and enable more detailed, specific and technical guidelines to be established by appropriate technical experts. The LTDP Common Guidelines document will be periodically updated to reflect the advances of activities carried out in the LTDP area and will be submitted, in the framework of the periodical updates, to public reviews to collect feedback and comments.

2.3 LTDP PRESERVE DATA SET CONTENT

LTDP Preserve Data Set Content (PDSC) indicates what to preserve in terms of data and associated knowledge and information during all mission phases to be able to satisfy the needs of the Earth Science Designed community today and in the future [5]. LTDP PDSC addresses the Earth Science context (i.e. Earth Science Domains) and the specific Earth Observation domain, based on the data categorization taxonomy described below.

Methods, standards and sound criteria are needed to certify whether the preserved data set content is complete and will meet the needs of future users. Long – term preservation requires solving to complementary yet coordinated problems:

- Preservation of the data records itself (the raw data bits acquired from an Earth Science instrument);
- Preservation of the context surrounding the data records (the meta-information needed to interpret the raw data bits).

The acceleration increase of the amount of digital information sensed by Earth Science instrument coupled with the aging of our existing digital heritage and well published examples of the impacts of its loss have raised the criticality and urgency of the sensed data record stream preservation.

In the frame of FIRST survey, [4] the user community has clearly and strongly pointed out that preserving data records of Earth science historical mission is mandatory. Particularly, the scientific community welcomes the LTDP European initiative to cooperate and optimize efforts to preserve data set heritage for future generation.

One of the outcomes of the FIRST study is that the criticality of preserving the context information is not a static attribute of the context information itself but a dynamic outcome of past commitments of the consumer community, information curators and holding institution. In the frame of FIRST a preliminary attempt has been made to rank context information criticality for a generic Earth Science mission and for the nine Earth science sensors types. This preliminary ranking will be tuned following the results of the pilot implementation projects initiated by the Agency to preserve ESA historical data set and their context information using the checklists as reference

In the frame of ESA survey, the user community has been also pointed out that non only data records but the latter context requires preservation too, as context information might often be:

- hidden or implicit: well understood in their respective designated user communities and data producer experts at the time the data records stream is acquired and processed;
- evolving the technological context (computer platforms, programming languages, applications, file format etc..) surrounding any piece of information will inevitably change over time until information is no longer usable; and the communities context (data producer, data consumer, designated communities i.e. communities and organization involved in the information's creation and initial use) may change over time and give different value to the data information over time of cease to exists.

The combination of the context surrounding earth science data information being both implicit and evolving requires that for the information to remain meaningful and useful after a long time span either the data records information must continuously evolve with the context, or the context must be captured and preserved along with the preservation of the data records, preferably at the time of the information creation.

The context surrounding earth science data records is particularly complex as stated in the introduction. For example use of remote sensing imagery requires detailed knowledge of sensor and platform characteristics, which due to its size and complexity is not usually bundled with data objects in the same way that the descriptive metadata is. Furthermore geospatial data may require deep analysis to remain usable over time.

As a major example, to support the long term climate change variables measurement, historical data records must be periodically reprocessed to conform to the most recent revisions of scientific understanding and modeling. This in turn requires access to and understanding of the original processing, including scientific papers, algorithm documentation, processing sources code, calibration tables and databases and ancillary datasets.

Whereas preservation of bits requires that the bits stay unchanged over time, preservation of context must accommodate and even embrace change to the context. File formats will need to be migrated over time, new access services will need to be developed and will require new kinds of support and information will inevitably be re-organized and re-contextualized.

Thus a key consideration in the design of an archive system is that it be able to capture and preserve complex contextual data records information objects; maintain persistent associations between data

records information objects and contextual objects; and support modification of the context over time.

2.3.1 Preservation Principles

The principles stated in the previous paragraph have been applied in the definition of the preserved data set content:

- minimum time reference for long term preservation is usually defined as the period of time exceeding the lifetime of the people, application and platforms that originally created the information;
- preservation of the data records (the raw data bits acquired from the mission instrument) is mandatory
- the data record context surrounding the information (hidden or implicit information) shall be addressed too when defining the preserved data set content;
- the context must be captured and preserved along with the data records, preferably at the time of the information creation and taking into account the evolving characteristics of context information, particularly for long term data series.
- the criticality of preserving the data set content is dynamic, an outcome of past commitments on the consumer community, information curators and holding institution

To analyse what shall be preserved, the approach is based on four main dimensions:

- Time dimension: How long? How long shall the data set content be preserved at minimum?
- The Content dimension referred to as What? What data set content shall be preserved?
- The Stage during which the dataset is generated When? When shall the information be captured and preserved?
- The past, current and future perceived importance (“persistency”) referred to as Rank: How critical is that the each information content object is preserved?

2.3.2 European EO Space Data Set

The European EO Space Data Set consists of all EO space data from missions or instruments owned by public or private organisations from European Member States and Canada and of all EO space data over Europe from non-European Member States missions or instruments available through agreements with European entities (e.g. Third Party Missions managed by the European Space Agency). The space missions or sensors whose data constitutes the European EO Space Data Set are subdivided in the following main categories:

- C1: High and Very High resolution SAR imaging missions/sensors (different Radar bands).
- C2: High and Very High resolution multi-spectral imaging missions/sensors.
- C3: Medium resolution Land and Ocean monitoring missions/sensors (e.g. wide swath ocean colour and surface temperature sensors, altimeter, etc.).
- C4: Atmospheric missions/sensors.
- C5: Other Scientific missions/sensors.

All missions and instruments comprising the European EO Space Data Set are described in a document, which is updated every six months.

3. ROADMAP VISION

ESA has developed the Standard Archive Format for Europe (SAFE) [6] an extension of the XFDU standard [7]. SAFE has been designed to act as a common format for archiving and conveying data within ESA Earth Observation archiving facilities. Many of the most important datasets have been converted to this format.

The important point is that XFDU, and therefore SAFE, is designed to implement the OAIS Archival Information Package [8], which in principle has everything needed for long term preservation of a piece of digitally encoded information. Some of the other components under consideration for the ESA LTDP technical implementation are the hardware needed to store the large volumes expected.

4. LTDP ARCHITECTURE

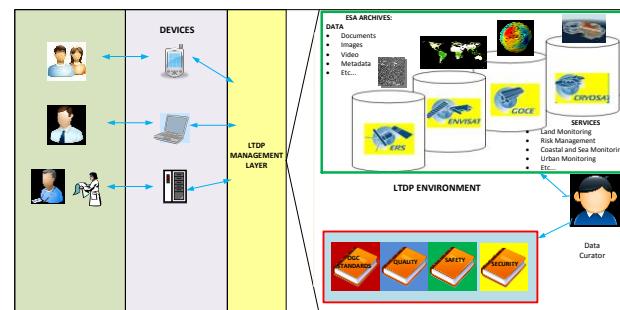


Figure 1 LTDP Architecture

Figure 1 shows the LTDP Architecture.

We distinguish different type of user, simple user or scientist, that with their devices access to LTDP Management Layer to obtain what they want. On the other side we have a Data Curator that is a crucial figure in LTDP Architecture. He preserves different mission data, documents, images, metadata, but particularly services with their technologies (i.e. Land Monitoring, Risk Management, etc....).

5. LTDP USE CASES

A use case expresses the functional, informational and qualitative requirements of a user (i.e. an actor or a stakeholder) whereby the functional requirements are represented by the „sequence of action“, and the informational requirements cover the content of the „observable result“. The qualitative needs encompass all the non-functional aspects of how the result is produced and the quality of the result which is important for the decision if the result is „of value“ to the user. Therefore, the degree of abstraction and formalism, and the language, should be such that it is adequate for the domain of expertise of the stakeholders. To serve as an agreement, it should be understandable to the stakeholders but also precise enough.

In this work, the concept of use cases is applied in order to describe the high-level functional requirements. We use the

Unified Modelling Language (UML) for this purpose, but by extending the UML use case notation with references to major information objects that are required to implement the use case.

Figure 2 shows the basic template that is used to present the LTDP use cases. Two major types of actors are distinguished: first, human and users that use a client application by means of its user interface; and second, „Software Component“ which represent a pieces of software that invokes an LTDP service by means of its service interface.

Use cases need information objects as inputs. These are indicated in the upper part of the diagram together with the required access method, i.e. create, read, write, delete. Results of use cases are listed as information objects in the lower part of the diagram. Information objects may be related to each other. Furthermore, use cases may have relationships to other use cases.

One use case may invoke another use case (which represents a dependency between use cases), or one use case may be a sub-variant of another.

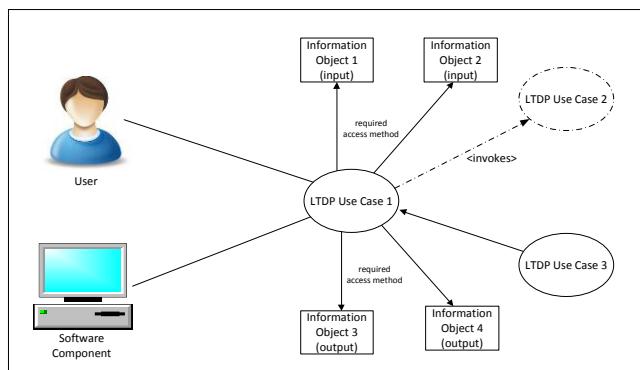


Figure 2 Main LTDP Use Case

5.1 DISCOVERY USE CASE

The Discovery Use Case deals with the question of how to find the EO resources (e.g. dataset, dataset series, services or sensors) of interest to a user. As in other application domain, such EO resources need to be described by some additional information, usually called metadata or metadata information. Metadata informs about the major characteristics of a resource. Metadata elements are stored in metadata stores (e.g. realised by relational databases) and accessed through interfaces of dedicated services.

The goal for the end user is to access those products that fulfil specific requirements according to his tasks. Essential requirements are, for instance:

- Region of interest
- Time series
- Usage of a specific satellite or sensor
- Corresponding ground station
- Additional attributes depending on the sensor type, e.g. cloud coverage.

As illustrated in figure 3, such requirements are entered as parameters in search queries. The access process delivers result sets that are specific to the resource types at which the search request has been targeted, i.e. delivers descriptions (metadata elements) of dataset series, sensors and /or services. The user may then browse through these metadata records and select those with which he wants to continue the interaction with other use cases.

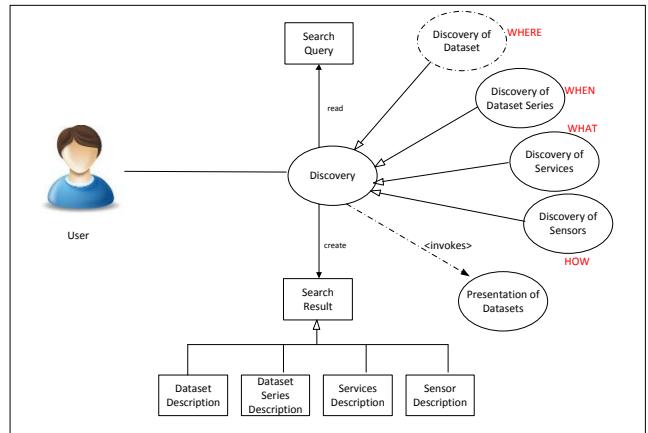


Figure 3 „Discovery“ Use Case

6. CONCLUSIONS AND FUTURE DEVELOPMENTS

The Future LTDP ESA Program targets the preservation of scientific data and associated knowledge from ESA and ESA-managed Third Party Missions in all fields of Space Science and in particular scientific data generated by payloads and instruments on-board space platforms (e.g. spacecraft, International Space Station). These activities have the following main objectives: ensure and secure the preservation of archived data and associated knowledge for an unlimited time span knowing that they represent a unique, valuable, independent and strategic resource owned by ESA Member States and ensure, enhance and facilitate archived data and associated knowledge accessibility through state of the art technology and exploitability by users, including reprocessing, for all the ESA and ESA-managed Third Party Missions in all fields of Space Science covered under the LTDP activities.

In cooperation with other space science data owners of Member States establish a cooperative, harmonized and shared approach to preserve and maintain accessibility to European Space Science Data for the long term.

7. REFERENCES

- [1] European Space Agency LTDP area on GSCB Website
<http://earth.esa.int/gscb/ltdp/objectivesLTDP.html>
- [2] European Space Agency LTDP Framework
<http://earth.esa.int/gscb/ltdp/EuropeanLTDPFramework.pdf>
- [3] European LTDP Common Guidelines, Draft
Version 2,
http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_Issue1.1.pdf
- [4] First LTDP Workshop, May 2008,
http://earth.esa.int/gscb/ltdp/LTDP_Agenda.html
- [5] European Space Agency –PDSC
<http://earth.esa.int/gscb/ltdp/EuropeanDataSetIssue1.0.pdf>
- [6] SAFE web site <http://earth.esa.int/SAFE/>

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

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

Requirements Elicitation for a Long Term Digital Preservation System: A Case Study from the Financial Sector

Claudia-Melania Chituc
University of Porto, Faculty of Engineering
Informatics Engineering Department
Portugal
cmchituc@fe.up.pt

Petra Ristau
JRC Capital Management Consultancy & Research
GmbH
Germany
+49-30-847 88 22-0
pristau@jrconline.com

ABSTRACT

Companies face challenges towards designing and implementing a preservation system to store the increasing amounts of digital data they produce and collect. The financial sector, in particular the investment business, is characterized by constantly increasing volumes of high frequency market and transaction data which need to be kept for long periods of time (e.g., due to regulatory compliance). Designing and developing a system ensuring long term preservation of digital data for this sector is a complex and difficult process. The work presented in this article has two main objectives: (1) to exhibit preservation challenges for the financial sector/ investment business, and (2) to present and discuss preliminary results of the requirements elicitation process, with focus on the financial sector - work pursued towards the design and development of a preservation system, within the scope of the on-going R&D FP7 project ENSURE – *Enabling kNowledge Sustainability Usability and Recovery for Economic value* (<http://ensure-fp7.eu>). Requirements, use cases and scenarios identified for the financial sector are presented and discussed. The agenda for future research work is also presented.

Categories and Subject Descriptors

H.3.4 [Information Storage and Retrieval]: Systems and Software

General Terms

Documentation, Design.

Keywords

Requirements elicitation, financial sector, digital preservation system.

1. INTRODUCTION

The financial domain, in particular the investment business, is characterized by the increasingly incoming streams of high frequency market data. The digital data produced and collected by financial institutions (e.g., market data, transactions data) needs to be preserved for long term (e.g., for regulatory compliance, research purposes). While in the last decade a particular focus of R&D in the financial business was on performance improvements of the IT-infrastructure in an attempt to keep pace with the constantly increasing volumes of data, today the need of financial institutes for support in compliance to regulations and legal standards takes an increasingly important role. In the light of the financial crisis, it can be well expected that the relevance of this issue will rise further, since various expansions of regulations are being discussed, like, for example, full transparency of transactions [1][2].

The design and development of a system ensuring long-term preservation of digital data is a complex and difficult process. Existing approaches focus mainly on preserving homogeneous data (e.g., cultural heritage data). In the case of the financial sector, the task to elaborate a preservation system is even more challenging since it is required to ensure long term usability of heterogeneous data, integrity and authenticity of significant intellectual property or highly personal data, while conforming to regulatory, contractual and legal requirements. Such challenges are addressed by the on-going R&D FP7 project ENSURE – *Enabling kNowledge Sustainability Usability and Recovery for Economic value* (<http://ensure-fp7.eu>) targeting the financial sector, health care, and clinical trials domains.

Once the necessity to elaborate a system ensuring long term preservation of digital data is expressed, the process of requirements elicitation starts. During the requirements collection and assessment phase, the properties (e.g., functional and non-functional properties) that the system should have to meet the stakeholder's needs are identified.

The work presented in this article has two main objectives: (1) to exhibit preservation challenges for the financial sector/ investment business, and (2) to present and discuss preliminary

results of the requirements elicitation process pursued towards the design and development of a preservation system, within the scope of the on-going R&D FP7 project ENSURE – *Enabling kNowledge Sustainability Usability and Recovery for Economic value* (<http://ensure-fp7.eu>). Relevant functional and non-functional requirements, use cases and scenarios identified for the financial sector are presented and discussed.

The rest of this article is organized as follows. The next section briefly introduces challenges in requirements elicitation in the context of long term digital preservation (LTDP). Section three describes the elicitation approach taken in the ENSURE FP7 project. Section four then portrays main characteristics of the financial sector and challenges towards long term digital preservation identified in this project. Relevant requirements, use cases and scenarios identified for the financial sector towards the design and development of the ENSURE Preservation System are then presented and discussed. The conclusions of the work pursued and directions of future research work are presented in section six.

2. LONG TERM DIGITAL PRESERVATION

2.1 Definition and Preconditions

Among the definitions found in literature, we exemplarily cite the following three. Digital preservation concerns the processes and activities related to preserving digital data over long periods of time, ensuring its accessibility and usability to future generations [3]. Digital preservation involves the retention of digital data/ object, and its meaning [4]. According to the Consultative Committee for Space Data Systems [5], long term preservation refers to the act of maintaining information, independently understandable by a specific community, supporting its authenticity over the long term.

The OAIS Reference Model (e.g., [5],[6]) presents a technical recommendation establishing a common framework of terms and concepts which make up an Open Archive Information System (OAIS). It comprises six functional entities (and related interfaces): Ingest, Archival Storage, Data Management, Administration, Preservation Planning, and Access.

Ensuring long term digital preservation (LTDP) is a complex process, and several challenges need to be addressed, such as: digital (technology) obsolescence, lack of standards and generally accepted methods for preserving digital information, deterioration (e.g., of digital data recording media). Although several approaches for digital preservation exist (e.g., emulation, information migration, encapsulation), as emphasized in [4], still, after ten years, there is a lack of proven preservation methods to ensure that the preserved digital data will (continue) to be readable after long periods of time, e.g., 20 years, 50 years, 100 years.

Although initially the focus has been on relatively homogeneous heritage data, currently organizations from the private sector (e.g., financial institutions, private clinics and hospitals) are increasingly concerned with preserving the growing amounts of digital data they produce and collect. These data tend to be heterogeneous, which represents an additional challenge for the

preservation process itself, but also for the elicitation of requirements.

2.2 Challenges in Requirements Elicitation

Commonly used approaches for requirements elicitation in the context of LTDP are: the questionnaire (e.g., CASPAR FP6 project [7]), and internal surveys and interviews (e.g., KEEP FP7 project [8]).

Although these methods for requirements elicitation allowed the identification of requirements and scenarios, they have several limitations, such as: lack of standardized procedures for structuring the information received through interviews, difficulty to integrate different answers/ interpretations, different goals, different communication styles or terminology into a single requirement.

In the context of long term digital preservation, several challenges appear in the requirements elicitation phase, such as: difficulty to validate the collected requirements (e.g., due to the very long life-time of a preservation system and preservation period), hardware and software constraints during system's design and implementation phases, changing requirements over time. The functionalities of a preservation system may also change over time. It is also very challenging to integrate into a single requirement (or a set of requirements) needs expressed by stakeholders from different industry sectors for one preservation system.

3. REQUIREMENTS ELICITATION FOR THE ENSURE PRESERVATION SYSTEM

3.1 The ENSURE FP7 Project

The on-going ENSURE FP7 project aims at extending the state-of-the-art in digital preservation by building a self-configuring software stack addressing both the configuration and preservation lifecycle processes in order to create a financially viable solution for a set of predefined requirements [9]. It analyzes the tradeoff between the costs of preservation against the value of preserved data, addressing also quality issues. ENSURE draws on actual use cases from health care, clinical trials, and financial services.

ENSURE Reference Architecture for long term digital preservation is based around four main areas of innovation¹: i) evaluating cost and value for different digital preservation solutions, ii) automation of the preservation lifecycle in a way which can integrate with organizations' existing workflow processes, iii) content-aware long term data protection to address privacy issues like new and changing regulations, and iv) obtaining a scalable solution by leveraging wider ICT innovations such as cloud technology.

3.2 Requirements Elicitation Approach

Considering the specificities of the LTDP domain, the objectives of the ENSURE FP7 project, and the characteristics of the three sectors on which this project is focusing on (financial, health-care and clinical trials), a use-case scenario approach has been

¹ <http://ensure-fp7.eu>

chosen for requirements elicitation, to reflect all the tasks the stakeholders will need to perform with the ENSURE preservation system. This was then combined with a traditional elicitation approach (e.g., where the stakeholders indicate what they want the system to do).

The use-case approach has been successfully used during the last years for requirements elicitation and modeling, e.g., [10], [11], [12]. As emphasized in [13], the objective of the use-case approach is to describe all the tasks the users/ stakeholders will need to perform with the system. Each use case can consist of several scenarios capturing user requirements by determining a sequence of interactions between the stakeholder and the system.

Within the scope of this work, use case scenarios were regarded as assets that can be exploited as reference scenarios within the context of the ENSURE project and the ENSURE Preservation System [1][2]. Thus, the description of the ENSURE use case scenarios focused on the expectations the technological solution (e.g., the ENSURE preservation system) should address. Similar to [10], these descriptions should bring out the goals (and assumptions) the technological solutions should encompass.

Although the use case approach is successfully used for capturing requirements (e.g., [10]), several weaknesses are documented, e.g., [14], such as: the use cases are written from the system's (and not stakeholders') point of view.

To avoid such pitfalls, a well-documented template for requirements elicitation was provided to the stakeholders (e.g., targeting functional and non-functional requirements, use cases and scenarios). In addition, sector-specific discussion groups/ teleconferences were set where experts in the area of LTDP participated, as well as representatives of the Architecture team.

The use of this approach allowed the identification of functional and non-functional requirements for the ENSURE preservation system, stakeholder's constraints for the technical solution, main classes of users, as well as the identification and documentation of use cases and scenarios.

4. THE FINANCIAL SECTOR

4.1 Main Characteristics and Challenges towards Long Term Preservation of Financial Digital Data

Before going into details about the data itself, we would like to address the sources and flows of the data as well as the stakeholders dealing with the preservation system. For the requirements elicitation process previously described, we restricted ourselves to the sector of investment banks and smaller financial investment institutes, whose core business is to offer advisory and asset management services to institutional as well as private investors.

So, one large source of data consists of the transaction documentation mainly in form of individual account statements from the custodian bank, broker or clearing house for each client. The constantly increasing amounts of documents go along with sharply increased requirements on information protection, safekeeping and risk management, and contain

expanded overall record retention obligations issued by the regulating public authorities at both national and European level. This concerns, for example, any information received from or provided to the client that needs to be preserved for the whole duration of the contractual relationship that could actually overpass the minimum statutory preservation time of typically 5 years.

On the other hand, almost all investment institutes base nowadays their advisory services on decision support systems for forecasting and algorithmic trading. These systems automatically analyze incoming real time data and come to trading suggestions when to buy and sell a specific instrument. Before actually being applied to real time data, the underlying trading models are developed and tested on the basis of long histories of market data, often reaching back for more than 5 to 10 years.

While there are no regulatory or legal requirements on the retention of these large volumes of high frequency *financial market data*, they have to be stored and retained, due to their associated business value, i.e. their high recovery and restocking costs [1][2].

Although each institute and situation is unique, there are common organizational characteristics shared by all investment institutes.

Figure 1 portrays the flows of data between the different departments of a typical investment bank:

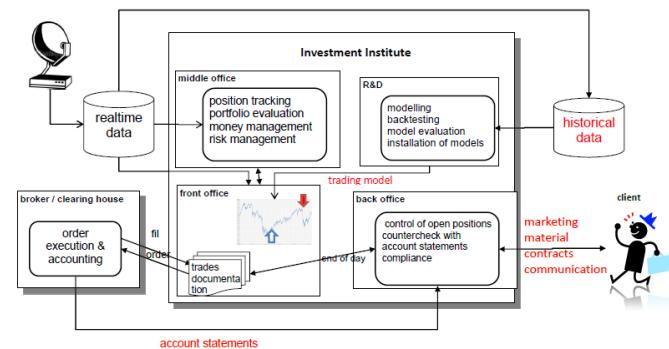


Figure 1. Flows of data between departments in an investment bank

Market Data is received from the real time feed and distributed to the historical data base from where it is accessed by the **financial engineers** located in the R&D department. Besides, it is of course immediately displayed on the charts on the traders' desktops in the front office as well as fed to the middle office, where it is used for tracking of open positions and risk management.

Client data, i.e. account statements, are received from the custodian bank, broker or clearing house. It is the **back office staff** that manages this type of data together with any other client related data and is hence responsible for the preservation of this type of data.

Trading Models, finally, are developed by the R&D department and installed at the trader's desks. So, the financial engineers are

the responsible actors for the retention of all models and market data.

A big challenge in the financial sector is the heterogeneity of the data that needs to be stored in the preservation system. For example, the *client information* consists of heterogeneous data, such as: contractual paper work, marketing information, digital information (e.g., e-mail correspondence, web content, electronic account statements, accompanying notes, like telephone protocols, front-office documentation about trading decisions and order execution).

Market data, on the other hand, may exist in heterogeneous formats as well. Technological developments on both data feed side but also concerning charting and modeling software typically lead to occasional switches between data and software providers. So the data format may vary between different proprietary formats from the various vendors as well as source and software independent formats, typically in comma separated ASCII files (*csv*), distributed by third party vendors of historical data.

Finally, the trading models themselves shall be preserved and build a third category of data. Although it is not a requirement yet, the preservation of software applications should also be considered since trading models critically depend on software versioning, e.g., in order to avoid eventual deviations in model interpretation.

Next are briefly described **data formats** for each category:

- *Client Documents* need to be kept in paper form (e.g., signed originals) and digitalized form (e.g., client contract, risk disclosure information and limited power of attorney will build the foundation of a client's record). Only standard formats will be used to preserve the data in the preservation system, e.g., *.pdf*, *.jpg*. In addition, some current MS Office document formats will have to be supported by the preservation system, such as protocols of telephone conversations.
- *Financial Market Data*. With the constant introduction of new financial trading instruments, but also observable in the well-established high volume instruments (e.g., Forex, Stock Index Futures like Dax and Nasdaq), the financial markets are characterized by extraordinary increasing data loads during the past years. Market data has to be distinguished by its sources (e.g., Thomas Reuters, Bloomberg, Morningstar), which use different data formats. It refers to numerical price data, reported from trading venues, such as stock exchange, but also from non-regulated interbank markets (i.e., foreign exchange rates). The price data is attached to a ticker symbol and additional data about the trade like stamp and volume information. The data is then stored in a database, typically one that is integrated with a charting environment software used by the trader to graphically display the data. Although the databases have their own proprietary format, almost all of them offer the possibility of exporting data to *csv* format.

A sample of financial market data in such a general format is illustrated in *Figure 2*. The column descriptors are contained in the first line (or in a separate file), and data is given in a ten minute compression format, i.e., all price data arriving within a 10 minute time window is reduced to 4 values only: the first price (Open), the highest (H); lowest (L) and last one (Close), the number of up-ticks (U, prices higher than the previous one), and down-ticks (D, prices less or equal to the previous one).

```
"Date","Time","O","H","L","C","U","D"
08/05/2009,1530,95.28,95.33,95.26,95.32,1528,1526
08/05/2009,1540,95.32,95.33,95.21,95.28,2520,2574
08/05/2009,1550,95.28,95.31,95.21,95.22,2634,2666
08/05/2009,1600,95.22,95.29,95.13,95.28,2936,2842
08/05/2009,1610,95.29,95.29,94.96,94.97,3930,3882
08/05/2009,1620,94.97,95.07,94.86,94.94,3710,3612
08/05/2009,1630,94.94,95.00,94.92,94.96,2946,2768
08/05/2009,1640,94.94,94.98,94.89,94.96,2526,2452
08/05/2009,1650,94.96,94.98,94.86,94.92,2618,2624
```

Figure 2. Sample of Financial Market Data
(Source: [1-2])

- *Trading Models* shall be preserved as code written within a proprietary programming language (e.g., depending on the modeling environment used), which typically contain some commands close to natural languages, allowing also trader with no or restricted programming skills to formulate their own trading rules. *Figure 3* presents a simple example of a trading model code snippet.

```
if Close crosses above Average(Close,10) then Alert;
If Marketposition=0 and Barssinceexit(1)> Barsflat and
    if Close < Close[1]
then Sell Short 10 contracts next bar at Market;
```

Figure 3. Example of Trading Model Data (Source: [1-2])

Challenges for the long term digital preservation of the financial data concern not only the heterogeneity of the data that needs to be preserved, but also the retention of the preserved information, e.g., the retention of client information, the retention of proprietary applications due to business purposes, and the retention of very large amounts of market data stored over time, while meeting regulatory directives and business goals.

The system ensuring the preservation of data for the financial sector needs also to allow conformance to regulatory, contractual and legal requirements, and management of long term authenticity and integrity of intellectual property and personal data.

With these considerations, an extensive work has been pursued to identify functional and non-functional requirements, use cases and scenarios for the financial sector for a digital preservation system.

4.2 Main Findings: Use Cases, Scenarios, Functional and Non-Functional Requirements

Main stakeholders for the ENSURE Preservation System, for the financial sector, include [1][2]:

- *Back-office Staff*, the primary user of the ENSURE Preservation System, is responsible for all administrative tasks related to the organization's clients.
- *Model Developer/ Financial Engineer* is the employee of the organization responsible for the implementation, development and testing of the trading models.
- *Auditor* is an actor external to the organization, who periodically checks the compliance to regulatory standards and completeness of documentation.
- *System Administrator* is the employee responsible for the technical management of the organization's system, e.g., installing and technically managing the ENSURE Preservation System in a specific context.

The UML use case representation of the ENSURE Preservation System functionality for the financial domain is illustrated in *Figure 4*. The interactions between the actors and the ENSURE Preservation System are indicated.

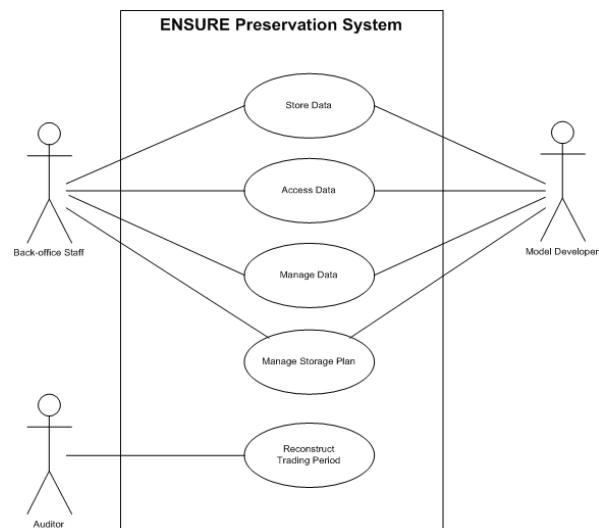


Figure 4. Main actors for the financial domain and the ENSURE Preservation System (UML) (Source: [1][2])

A brief description of the use cases illustrated in *Figure 4* is presented in *Table 1*.

Table 1. Overview of the Use Cases

Use Case	Description
Store Data	Data that is not accessed regularly is being preserved due to regulatory requirements or because it represents a business value.
Access Data	The data being preserved is accessed by the Back-office Staff or Financial Engineer.
Manage Data	Accompanying descriptions or metadata related to existing stored data of any type may have changed and needs to be updated, e.g., a client contract may have terminated or a trading model may no longer be in use. It is important to update these changes; especially in the case of client documents the safekeeping period starts with the termination date.
Manage Storage Plan	A storage plan describing how and where to store which type of data is set up once and used for each storage/ access operation. With respect to the available storage hierarchy and considering the related costs and risks of each device, the system determines the optimal storage location for each type of data and makes a suggestion to the user. The user may agree to the suggested plan or alter it according to additional constraints. The plan remains then fixed until it is updated on user request, e.g., when new storage devices are available or when related costs have changed.
Audit documents	A sample of client documents is requested (e.g. client XY for the period 01/2011 – 12/2011) by the auditor. It is then checked for correctness and in particular for completeness.

Table 2 contains a brief description of the *Store Market Data* scenario for the financial sector.

Table 2. Store Market Data Scenario

(Source: adapted after [1] and [2])

Name of scenario	Store market data
Actors	Financial Engineer, System Administrator
Operational Description	Once a month market data for all markets received by the real time data feed will be stored in the ENSURE Preservation System.
Problems, Challenges	Data integrity. Data protection. Large amounts of input data in one chunk.
Expected Benefits	Prevent real time database from overload. Secure business value of data.
Risks	Losing or corrupting data. Degradation of system performance during ingest of data.

Figure 5 illustrates the UML sequence diagram for the Back-office Staff for a simple scenario on client data request. After

the *Back-office staff* submits the login and password, the ENSURE preservation system shall send a success/failure notification. A successful login allows the *Back-office staff* to submit requests (e.g., query on client data, query on client name). The ENSURE preservation system shall verify the access rights for the query received and return the data requested.

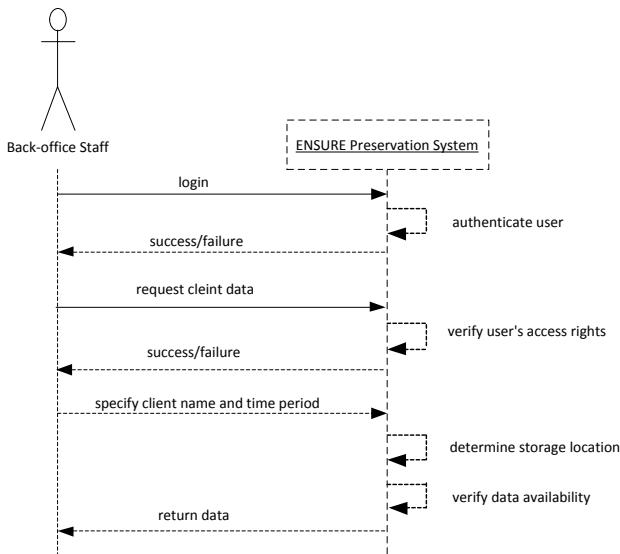


Figure 5. Client data request sequence diagram for the ENSURE Preservation System (UML)

The approach used also allowed the collection of functional and non-functional requirements. Five relevant *functional requirements* for the ENSURE preservation system identified for the financial sector are listed below [1][2]:

- *Authenticate User.* The ENSURE preservation system shall ensure that only authorized users have access to the system and are restricted to executing only those tasks necessary within their specific assigned role.
- *Encrypt Data.* The financial data is subject to security restrictions and shall be encrypted before being stored on an external storage device. The ENSURE preservation system shall detect automatically which data has to be encrypted and shall start the encryption process.
- *Notify Registration.* When a user is registered by the System Administrator, the ENSURE preservation system shall send an e-mail to the user with his/ her access details.
- *Keep Track of Minimum Regulatory Storage Times.* The ENSURE preservation system shall attach to each record of client documents the minimum storage duration. This requires also the possibility of entering the date when a client relationship has ended. The ENSURE preservation system shall automatically label the data with the earliest possible deletion date.
- *Delete Old Client Documents.* Client documents that have reached the minimum storage duration shall be

deleted in order to relieve the system from unnecessary volume. The ENSURE preservation system shall detect such data automatically and ask for user confirmation before deleting the data.

- *Evaluate and optimize a preservation solution* considering cost, quality and economic performance in order to support differently experienced users.
- *Allow for the Reproduction of Results of Trading Models* (which is version sensitive).

Examples of *non-functional requirements* for the ENSURE preservation system identified for the financial sector are [1][2]:

- *Availability.* The ENSURE preservation system shall be available every working day during normal working hours (e.g., 9h-18h). Since the storage of data will be done in weekly or monthly periods, and retrieval will be executed without time constraints, non-availability of the system for less than 24 hours is not critical.
- *Reliability.* Resilience and reliability of the system are critical factors with respect to compliance to regulations, e.g., loss of data is not acceptable.
- *Integrity.* Data integrity is very important due to compliance to regulatory prescriptions (e.g., changes or partial loss of client data is not acceptable), except for market data.
- *Confidentiality,* which concerns client documents and trading models. Stored trading models contain proprietary know-how and the essence and experience of many years of research. For this reason, the integrated trading models represent a high economic value for a financial institution that has to be protected against unauthorized use and reverse engineering attacks. Furthermore, in order to ensure the correct restoration of trading signals, the models have to be protected against external manipulation and corruption. The sensitive nature and strategic value of both such data cannot tolerate any unauthorized access or usage.

5. DISCUSSION

The data considered for long term preservation in the financial use case within the ENSURE project consists of client data (including transaction documentation as, for example, daily account statements) on the one hand, that has to be stored due to record retention obligations according to legal regulations like the EU's Markets in Financial Instruments Directive² (MiFID), and market price data histories on the other hand needed by researchers and market analysts in order to build prognosis models to be used by the traders as decision support systems. As different as the goals for data preservation – legal reasons for client data and economic reasons due to high costs for

²http://europa.eu/legislation_summaries/internal_market/single_market_services/financial_services_general_framework/l24036_en.htm

repurchasing, filtering and format conversions for market data – as different are the consequences of data loss and security violations on both data classes.

Since the legal consequences of losing client data can be severe, up to endangering the continuance of business execution, the preservation plan has to fulfill highest service level with respect to security standards including duplication of data and multiple fixity checks. For the same reason, cost considerations will only come in the second place for this type of data, after all data security requirements have been met. For market data, in contrast, storage costs represent the main decision criteria during preservation plan generation.

Another major criterion for the preservation strategy stems from the different data formats found in both data classes. While client data is typically stored in widespread document types like *pdf*-s, market data format is determined by the origin of the data (i.e., the data feed provider) and/or the proprietary data format of the data base, where the data is fed to. While the handling of such common data formats like *pdf*-files does not represent a major challenge to a preservation system, proprietary data formats may call for two options to be followed: they may either be converted to more common or “standard” formats, independent from the original application software, or a virtualization path may be followed, preserving the whole application package, including data base, and making them accessible through the use of virtual machines.

A particular challenge consists of the fact that, as far as market data is concerned, its value decreases with its age. The target preservation plan shall therefore distinguish between relatively *recent data*, that has to be kept as the complete data set, so called *tick data*, consisting of every single price change, leading to hundreds of prices per minute, and *older data*, where compressed data format (so called OHLC³ bar representation) would be acceptable, reducing the data to only four values per time period. The resulting information loss would be acceptable, as well as a downgrade of the service level.

Following the path indicated in [4] for the selection of preservation techniques, all of the above considerations lead to a twofold strategy. Although the complexity of the digital resource is low for both data classes, the data format is not known as far as market data is concerned. In this case, an emulation solution is recommended for market data preservation while encapsulation or migration would be the technique of choice regarding client data.

6. CONCLUSIONS AND FUTURE WORK

Requirements elicitation for a preservation system is a complex task, mainly due to the challenges associated to long term digital preservation, such as: difficulty to validate and test a preservation system due to its very long lifetime, the data and technology to be preserved are exposed to obsolescence.

So far, most approaches towards ensuring long term preservation of digital data focused on relatively homogeneous data (e.g.,

cultural heritage data). In this article, challenges for the financial sector towards the design and development of the ENSURE preservation system were identified and discussed, e.g., data heterogeneity, conformance to regulatory, contractual and legal requirements, integrity of preserved data.

The ENSURE preservation system aims at providing support to perform an analysis of cost, economic performance assessment and quality for a preservation solution, which represents a novelty compared to other initiatives (e.g., iRODS⁴ policy-based data management system).

Results of the requirements elicitation process, with focus on the financial sector have been presented (e.g., functional and non-functional requirements, use cases and scenarios) which reflect the work pursued towards the design and development of a preservation system, within the scope of the on-going R&D FP7 project ENSURE: *Enabling kNowledge Sustainability Usability and Recovery for Economic value* (<http://ensure-fp7.eu>). The approach used in requirements elicitation for the ENSURE preservation system was also described.

The main result of the analysis presented in this article concerning the financial sector was that due to the heterogeneous character of the data to be stored a combined strategy will most probably be the most suitable one for the analyzed data. How exactly the strategy should look like and whether it is economically viable to use several storage devices in parallel will be the result of the next step in the ENSURE project. With the help of the ENSURE economic performance assessment engine (e.g., [15], [16]), several storage scenarios can be compared and their expected gains will be estimated and considered with respect to the given constraints for the data. Future work will also focus on the validation of the functional and non-functional requirements, use cases and scenarios identified, and their traceability in the implemented preservation system.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under the grant no. 270000. The authors acknowledge all consortium members who contributed to the ENSURE M4 and M12 Requirements Deliverables: D1.2.1a and D1.2.1b.

8. REFERENCES

- [1] ENSURE: D1.2.1a, ENSURE Requirements Deliverable, M4, June 2011 (<http://ensure-fp7.eu>).
- [2] ENSURE: D1.2.1b, ENSURE Requirements Deliverable, M12, January 2012 (<http://ensure-fp7.eu>).
- [3] Borghoff, U.M.; Rodig, P.; J. Scheffczyk; Schmitz, L. Long-term Preservation of Digital Documents, Springer Berlin Heidelberg New York, 2003.
- [4] Lee, K.-O; Stattery, O.; Lu, R.; Tang X.; McCrary, V. The State of the Art and Practice in Digital Preservation”.

³ Open-High-Low-Close.

⁴ <http://www.irods.org>

- Journal of Research of the National Institute of Standards and Technology, 107, 2002, pp. 93-106.
- [5] CCSDS – The Consultative Committee for Space Data Systems, “Reference Model for an Open Archival Information System (OAIS)”, Draft Recommended Standard, CCSDS 650.0-P-1.1, Pink Book, August 2009. (Accessed at: <http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%20650P11/Attachments/650x0p11.pdf> on February 20, 2012).
 - [6] Giaretta, D. Advanced Digital Preservation, Springer Verlag Berlin Heidelberg, 2011.
 - [7] CASPAR FP6 Project, D4101 User Requirements and Scenario Specification, 2006.
 - [8] KEEF FP7 Project, D2.2a Requirements and design documents for services and architecture of emulation framework, Part A, 2010.
 - [9] Edelstein, O., Factor, M., King, R., Risso, T., Salant, E. and Taylor, P., “Evolving Domains, problems and Solutions for Long Term Digital Preservation”. Proceedings of iPRES 2011, November 1-4, 2011, Singapore, pp. 194-204.
 - [10] DiNetto E., Plebani P., “Describing case-studies: the S-Cube approach”, 2010, S-Cube Networks of Excellence, <http://www.ss-cube-network.eu> [Accessed at: <http://www.ss-cube-network.eu/pdfs/EC-document-20100713.pdf> on February 3, 2011.]
 - [11] Cockburn, A. “Writing effective use cases”, Boston, MA: Addison-Wesley, 1999.
 - [12] Jacobson, M. “Software requirements & specifications: a lexicon of practice, principles and prejudices”. New York, NY, USA: ACM Press: Addison-Wesley Publishing Co., 1995.
 - [13] Wiegert, K.E. “Software Requirements”, Microsoft Press, 2nd Edition, 2003.
 - [14] Lilly, S. “Use case pitfalls: top 10 problems from real projects using use cases”. TOOLS 30, Proceedings of Technology of Object-Oriented Languages and Systems, 1999, pp. 174-183.
 - [15] ENSURE: D20.1.1, ENSURE Activity II Scientific Report, February 2012.
 - [16] Chituc C.-M., Ristau P. “A service-oriented approach to assess the value of digital preservation”. Submitted at the 2nd International Workshop on Performance Assessment and Auditing in Service Computing-PAASC2012, to be held at ICSOC 2012, November 12-16, 2012, Shanghai (*under review*).

Web Archiving Effort in National Library of China

Qu Yunpeng
National Library of China

No.33 Zhongguancun Nandajie, Haidian
District
Beijing, China

quyp@nlc.gov.cn

ABSTRACT

In this paper we introduce the effort in National Library of China in recent years, including resources accumulation, software development and works in Promotion Project in China. We have developed a platform for Chinese web archiving. And we are building some sites to propagate our works to the nation. At last we figure out some questions about the web archiving in China.

Categories and Subject Descriptors

<http://www.acm.org/class/1998/>

General Terms

Keywords

National Library of China, Web archiving

1. INTRODUCTION

Nowadays the web resources in Chinese language are growing in a very fast speed. According to the <29th China Internet Development Statistics Report> from CNNIC, up to Dec. 2011, the total number of sites in China reached 2.3 million. The number of pages reached 88.6 billion, with the Annual growth rate of over 40%.^[1]

However, the web resources are also disappearing rapidly. The Chinese web resources are the civilization achievement of Chinese people and the important part of the digital heritage of Chinese culture. They need to be preserved and protected. In 2003, WICP (Web Information Collection and Preservation) Project was found and some experiments were done. After a few years of researches and tests, we made some progress in Web Archiving.

2. Related Research

In the 1990s, web information archiving was focused by some institutions and organizations. The libraries, archives, research institutes started to do research and experiments on Web Archiving. The national library all over the world took Web Archiving as their duty-bound mission. In the Europe and America, some national libraries found their Web Archiving projects, for example, the Mineval^[2] project of Library of Congress, the Pandora of National Library of Australia^[3], the Kulturarv project of National Library of Sweden^[4], and so on. They accumulated much valuable experience for us.

Up to now, the preservation of Chinese web resources are still in the stage of Theoretical research and testing phase. In many colleges and research institutes, digital preservation is carried as an issue. There are two main testing projects for Web Archiving in China. One is the Web Infomall of Peking University. The other is the WICP (Web Information Collection and Preservation) project of National Library of China. The Web Infomall is carried

by the Computer Networks and Distributed Systems Laboratory of Peking University, with the support of national '973' and 985 projects. It is a Chinese web history storage and access system. It collected the Chinese web resources from 2002 with the amount of 3 billion pages, and now going with the speed of 45 million pages per day. In the site of Web Infomall, users can view the past pages, and view the selected events pages^[5]. In 2003, WICP in National Library of China was started to preserve the web resources.

3. The WICP Project

3.1 The Introduction

In the early 2003, WICP project was found and a team was established. The team consisted of the reference librarians, cataloguers and network managers. 4 problems needed to be solved by the team

- i. To find the problems in the collection, integration, catalogue, preservation and access of the web resources, and find the answers to them
- ii. Experimentally, to harvest those web information that can reflect the development progress of the politics, culture, finance and society of our country, and provide access to those long-term preserved.
- iii. To find the objects, the policy and the measures of the Web Archiving in NLC (National Library of China), so the technical routes and policies can be made accordingly.
- iv. To find the scheme of the business and to promote the integration of the web archiving business.

In the early stage, the main jobs are the researches and the software testing. In 2005, with the cooperation with Peking University, we preserved 19968 governmental sites which were registered in China and the domain name ended with 'gov.cn'. From 2006 we start to collect by ourselves. Up to now, we have a collection of web resources about 20TB, including 70 events from 2542 sites and 80000 government sites*harvest.

In 2009, we put the site 'China Events' on web and it can be accessed through internet. 'China Events' are based on the events crawling and preservation mentioned previously mentioned. 'China Events' are organized by the important historical events, selecting the news from archived resources and form multi-events contents. Users can search the metadata and browse the archived web sites. The events in 2008 consist of 10 events, such as the southern snow damage, the National People's Congress and Chinese People's Political Consultative Conference, the 5.12 Wenchuan Earthquake, the Olympics in Beijing, the Paralympics in Beijing, the launching of Shenzhou VII spaceship and so on.

The events in 2007 consist of 8 events including the 85th anniversary for Communist Party, the 10th anniversary for the return of Hong Kong., the Special Olympics in Shanghai, the 17th CPC National Congress, and Chang'e-1 lunar probe and so on. The events in 2006 include 7 events. They are Construction of new Rural, the Eleventh Five Year Plan, the 2006 International Culture Industries Fair, the 70th anniversary of Long-March, the Opening of the Qinghai-Tibet Railway and so on.

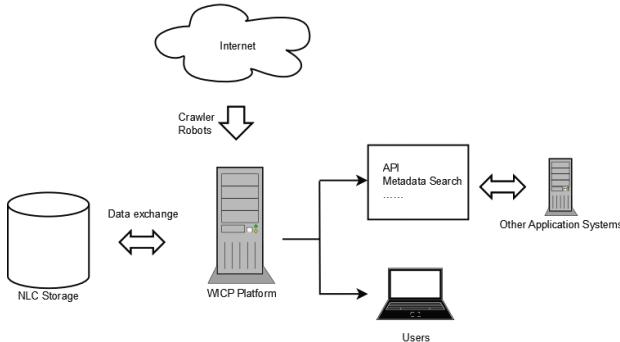


Figure 1. the framework of WICP

Figure 1 shows the framework of WICP. We collect resources from the internet using crawlers and robots, and put the preserved data to the storage. Other systems can use the data and exchange with WICP by API and other methods.

3.2 The Harvesting Policies

Comparing to the traditional information resources, web resources are tremendous, wide-spread and quickly increased. Also they are free to publish, and come from all kinds of sources. So they have the characteristics of complication and unevenness. So we consider that we need not to preserve them all. So according to the function of NLC being the repository of the nation's publications and a comprehensive research library, after a adequate research, we decide to use 'full harvest' policy for government sites mirroring and 'selecting harvest' for news preserving. In the ordinance of literature selecting of NLC, it wrote that the selecting of web resources should be done by event; the great or important events about the politics, finance, culture, sports, science and education should be focused.

4. The Distributed Harvesting Platform

4.1 The Motivation

In 2006, the focus of the project was turned to the technical problems in web archiving. After the comparing between the well-known sites and testing on the open source software, we decide to use the software that IIPC provide, including Heritrix, Nutchwax and Wayback.

After a few years of using, we find it inconvenient. There are some points:

- i. The guiding documents and the language are written in English, it is not easy for Chinese people to understand the exact meaning.
- ii. The open source softwares have their own functions, Heritrix for crawling, Nutchwax for full-text indexing and Wayback for url indexing and accessing. So if we want to finish the whole task, we should switch between softwares from time to time. Especially when we have several servers, it is a annoying job to handle all the stuff in these servers.

- iii. The analyzer in Lucene in Nutchwax does not perform good for Chinese language.
- iv. Some jobs have to be done outside the software, such as the cataloguing, the authorization for crawling, the statistics, and so on.

So, to solve these questions, we start to design a framework for a integrated platform. It covers all the function that heritrix, nutchwax and wayback have, and make them a smooth workflow including cataloguing and statistics. It can manage the servers in a distributed environment so that we do not have to change from one server to another from time to time. Its UI are in Chinese, supporting multi-language switch. The framework is show in figure2.

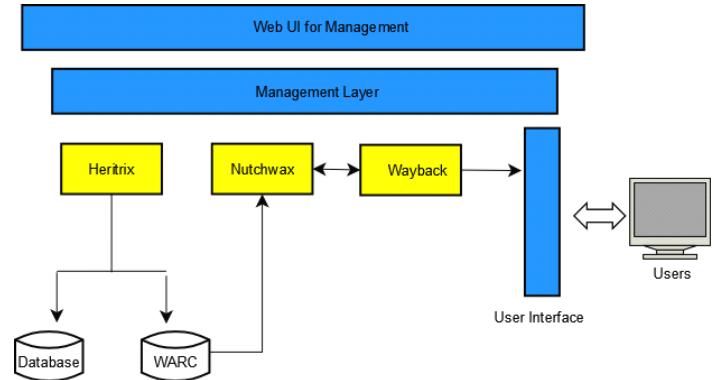


Figure 2. The framework of the Platform

4.2 The framework

After talking to some experts in computer and networks, we decided to put the platform on a distributed storage, for the performance of the platform and easy extension of the space. The Figure 3 is the primary design of the platform. The central controller is the kernel of the platform. Several crawlers are connected to the controller, saving the WARC to the Distributed storage through the controller. The specific information about tasks and crawlers are saved to a database. After indexing, users can access to the pages by the controller.

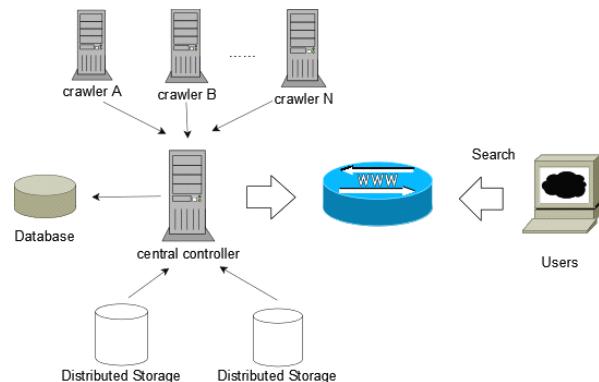


Figure 3. design of the platform

We do research on the open source softwares, and finally decided to program our platform based on Web Curator Tool^[6]. It is an open-source workflow management application for selective web archiving. It is designed for use in libraries and other collecting organizations, and supports collection by non-technical users while still allowing complete control of the web harvesting process. It is integrated with the Heritrix web crawler and

supports key processes such as permissions, job scheduling, harvesting, quality review, and the collection of descriptive metadata. WCT was developed in 2006 as a collaborative effort by the National Library of New Zealand and the British Library, initiated by the International Internet Preservation Consortium. From version 1.3 WCT software is maintained by Oakleigh Consulting Ltd, under contract to the British Library. WCT is available under the terms of the Apache Public License.

However, Web curator tool did not support management of multiple crawlers, and did not run on a distributed storage. So we need make some changes.

First, we must connect the crawlers to the controllers, so that the controller can check the status of the crawlers and assign tasks to them. In the implementation we use socket to connect the controller and the crawlers.

Second, we need to build a distributed storage and put the controller on them. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures [7]. So we decide to use Hadoop as the distributed storage environment. Figure 4 is the implementation of the platform.

Third, we need change the analyzer to support a good performance on Chinese.

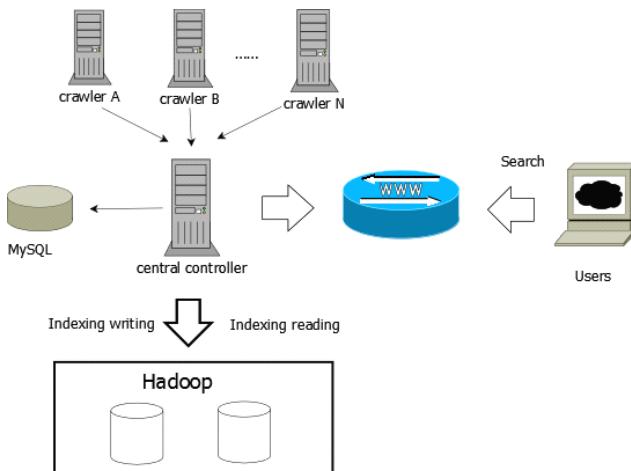


Figure 4. Implementation of the platform

4.3 Functions

Now the platform is under development, but the functions are ok.

It has all the functions WCT have, such as permission, harvesting, configuration and so on. Figure 5 is the main page of the platform. There is a new module in the UI, the cataloguing, which page is in the figure 6.



Figure 5. Management UI

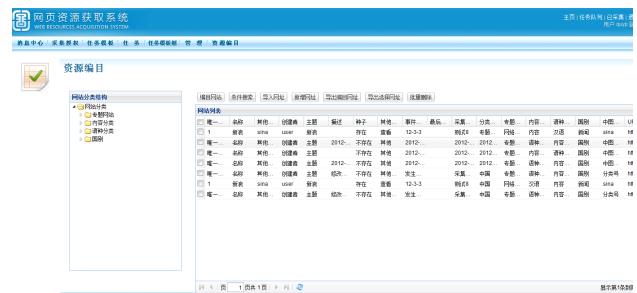


Figure 6. The Cataloguing module

NLC has developed a draft for internet items cataloguing, but it is still trying by experts. In this platform we only adopt several core elements in DC, because the key point of the platform is not on cataloguing. The platform provide a template of excel file. Librarians using the platform can fill in the excel file the metadata and submit to the platform once.

In the Configuration Module we can see the status of the multiple crawlers. When making the target, librarian can choose which crawler to run the task. If not, after 5minutes the system will assign a crawler to run automatically.

5. The promotion Project

The Promotion Project of Digital Library was established by the Ministry of Culture, Ministry of Finance. NLC is main force of the project. The project will apply VPN, digital resources, servers and application systems to the public libraries all over the nation, provincial, municipal and county level.

The platform was included in the application systems of the Promotion Project, as an optional one. This is a good opportunity for us to propagate the web archiving idea and our progress in this field. We have done some effort for the project to propagate itself nationwide. We are building a site for the knowledge of the web archiving and rebuilding the site of 'China Events'. In a few months we will give a training session in the conference of the Promotion Project, for the platform to the librarians all over China.

Web archiving is not an easy job for the public libraries in china. It needs great investment. It needs high- power servers, large capacity storage, enough bandwidth, and many human forces. The Promotion Project solves some of the financial problems, so this is a good opportunity for us to make web archiving understood, accepted and tried. The platform promotion is the first step of it. Next we will organize the libraries and archives together to do web archiving.

6. Conclusion and Future Works

In this article we present the progress of the web archiving effort in National Library of China. We have started the project for about 10 years, and have accumulate 20TB archived resources and much experience. In recent years we made several step, such as the software platform and works in Promotion Projects. But there is still long way to go, both for NLC and for Web archiving in China.

6.1 Legal Problems

The legal problem is the first obstacle that libraries meet when they are harvesting, preserving and using web resources. There are many conflicts against the existing copyright law. In order to solve this problem, many countries permit the deposit of web resources by legislation, such as Denmark, Sweden [8], France, Finland, etc. But in China, the copyrights and existing deposit regulations do not cover the web resources. And there is a blank for deposit of internet publications. That means there is no special laws and regulations for deposit of web resources. In order to form the national deposit system, the web resources should included in the deposit range of the ‘Chinese Library Law’. The web resources could be preserved completely by the law.

6.2 Cooperation

Web archiving action involves multiple factors, including policies, laws, finance, technique and management. It is large-scaled, heavy-invested, complex and persistent. Single institute could not take the heavy response and take the hard job. So we need to coordinate all the society resources by the means of cooperation. The starting stage of the web resources preservation lacks the unified planning and no institute is specified to take the responsible to preserve the web resources. So the situation is, some resources are collected by different institutes for several times and human power and money are wasted. Meanwhile, large amount of web resources are left unprotected. In many countries, national libraries are the main force to preserve and archive the web resources. They bring us a lot of references and Inspirations. NLC now are trying to form a cooperation system for the preservation and archiving of web resources, and those are mentioned above.

7. References

- [1] 29th China Internet Development Statistics Report [EB /OL], [2012-06-11].
<http://www.cnnic.net.cn/dtygg/dtgg/201201/W020120116337628870651.pdf>
- [2] Library of Congress Web Archives Minerva [EB /OL].[2012-06-11]. <http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html>
- [3] PANDOR [EB /OL]. [2012-06-11].
<http://pandora.nla.gov.au/>
- [4] The Kulturarw3 Project — the Swedish Royal Web Archive [EB /OL].[2009-03-01].
<http://www.emeraldinsight.com/Insight/ViewContentServlet?contentType=Article&filename=/published/emeraldfulltext/article/pdf/2630160205.pdf>
- [5] Web InfoMall [EB /OL].[2009-03 -02].
<http://www.infomall.cn/>
- [6] Web Curator Tool[EB /OL].[2009-03 -02],
<http://webcurator.sourceforge.net/>
- [7] Apache hadoop[EB /OL].[2009-03 -02],
<http://hadoop.apache.org/>
- [8] Zhong Changqing, Yang Daoling. Legal Issue in Web Resource Preservation [J].Information studies: theory and application, 2006 (3) : 281 – 284
- [9] Wang Ting, Wu Zhenxin, Gao Fan. Analysis of the International Collaboration Mechanism of the Preservation of Network Information Resources [J].Library Development, 2009 (3) : 6 – 13

Addressing data management training needs: a practice-based approach from the UK

Laura Molloy

Humanities Advanced Technology and
Information Institute (HATII)
University of Glasgow
Glasgow G12 8QJ
+44 (0)141 330 7133

Laura.Molloy@glasgow.ac.uk

Simon Hodson

Joint Information Systems
Committee (JISC)
5 Lancaster Place
London WC2E 7EN
+44 (0) 203 006 6071

S.Hodson@jisc.ac.uk

Stéphane Goldstein

Research Information Network (RIN)
20-24 Tavistock Sq
London WC1H 9HF
+44 (0)20 3397 3647

Stephane.Goldstein@researchinfonet.org

Joy Davidson

Humanities Advanced Technology and
Information Institute (HATII)
University of Glasgow
Glasgow G12 8QJ
+44 (0)141 330 8592

Joy.Davidson@glasgow.ac.uk

ABSTRACT

In this paper, we describe the current challenges to the effective management and preservation of research data in UK universities, and the response provided by the JISC Managing Research Data programme.

This paper will discuss, *inter alia*, the findings and conclusions from data management training projects of the first iteration of the programme and how they informed the design of the second, paying particular attention to initiatives to develop and embed training materials.

Keywords

Research data management; training; skills; digital curation; digital preservation; universities; research infrastructure; research support staff; postgraduate student research training.

1. INTRODUCTION: THE RESEARCH DATA MANAGEMENT CHALLENGE

The effective management of research data is an integral and inseparable part of the research process. Good research data management (RDM) therefore equates with sound research, a view which is reiterated by the Research Councils UK (RCUK) common principles on data policy [1] and the recent introduction

by the UK's Engineering and Physical Sciences Research Council (EPSRC)'s explicit expectations for the management of research data generated by funded projects. Many other initiatives echo this view, such as the Organisation for Economic Co-operation and Development (OECD) principles for access to research data from publicly-funded research [2] and the ongoing efforts by Neelie Kroes, Vice-President of the European Commission responsible for the Digital Agenda for Europe¹ including the All European Academies (ALLEA) declaration on open science of April 2012 [3].

For the purposes of our discussion here, we are using the term 'data management' broadly to incorporate the notions of digital curation² and digital preservation³, both as applied to research data produced by universities and other research institutions.

The challenge of achieving better RDM does not simply rely on addressing technical issues. These are important but tractable; equally important are organisational, policy and attitudinal issues. Universities are developing policies and technical infrastructure, but ultimately, researchers themselves have to be aware of the need for research data management, recognise that they have a role in managing their data, be willing to engage in RDM practice and have the skills, incentives and support to do so. Changes to the way research data is managed imply cultural change in the way research is practiced, whilst also continuing to support robust research processes. Disciplines vary in their levels of existing

¹Described at http://ec.europa.eu/information_society/digital-agenda/index_en.htm

² See the Digital Curation Centre's definition at <http://www.dcc.ac.uk/digital-curation/what-digital-curation>

³ See the Digital Preservation Coalition's definition at <http://www.dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts>

awareness and in common practice in the management and sharing of research data, so that researchers' behaviour is strongly influenced by their immediate environment as well as their disciplinary culture.

Researcher behaviour is specifically influenced by funders' requirements and increased recognition of the benefits of data sharing and re-use. Increased awareness in the researcher population of the options and advantages of data management and sharing can enable researchers to participate more fully in the emerging digital research economy.

The emerging digital research economy, as examined by the Royal Society's 'Science as a Public Enterprise' initiative, culminating in the 'Science as an Open Enterprise' report [4], has the potential to lead to significant advances in research and improve its overall quality by the provision of easier verification or reproducibility of the data underlying research publications. This can in turn allow new research questions to be asked of existing data, or integration of multiple datasets to achieve wider or more robust research conclusions. Research funders are naturally keen to obtain the greatest possible return on investment for grants disbursed. Increasingly, this is accepted to mean ensuring that research data is available for reuse and repurposing. Where the data is the product of unrepeatable observations, the case is easy to make. A well-known example is that there have been more papers published based on the reuse of archived data from the Hubble Space Telescope than those based on the use originally described when specific observations were requested.⁴ Funders are increasingly aware that the potential of reuse can be extended to other research areas: as the research data management principles developed by the EPSRC state, sharing and promoting the reuse of research data is an important contributor to the impact of publicly funded research [5].

Journal editors are also sensitive to the need for research data to be available for verification and reuse. A growing number of journals are adopting increasingly stringent data availability policies. Most innovative and significant among these, perhaps, is the Joint Data Archiving Policy which underpins the Dryad Data Archive initiative [6]. Since August 2011, twenty BioMed Central titles have adopted data availability policies of varying rigour. Enthusiasm is growing around the idea of data papers and of data publications – or at the very least more effective linking and visualization of data through traditional publications. As the Opportunities for Data Exchange (ODE) Report on Linking Data and Publications [7] testifies, there are a growing number of innovative initiatives and in the next few years the publication of research data is likely to be recognized as a necessary part of the

publication of research results.⁵ This has implications for the way in which researchers are trained.

Currently, a lot of publicly-funded data generated by UK universities is lost or inaccessible: this can cause serious difficulties for the researcher in the future when trying to re-access their data, and also greatly limits the possible return on the initial investment in that research. More sophisticated management of research data and improved linkage between research data and published outputs, then, clearly allows original research activity to be further exploited, yielding richer knowledge and wider impact. At the institutional level, there is currently a realisation in many universities that a significant change has to take place if these risks are to be controlled and these benefits are to be achieved.

2. RESPONDING TO THE CHALLENGES

In response to this set of challenges and the emerging new landscape within UK research, the JISC established the Managing Research Data (MRD) programme which has run as an ongoing set of activities since 2009. Through two iterations, the basic structure of the programme remains the same. Projects have tackled 'hard' (i.e. technical) and 'soft' (i.e. human) infrastructure challenges from the point of view of particular disciplines and, increasingly, specific institutions.

The work of the programme has addressed the practical and technical issues associated with research data management infrastructure and the challenges of data management planning, data citation, description, linking and publication. There has also been attention paid to the importance of training requirements. The first iteration of the programme (2009-11) funded five projects to address the training aspect which were supplemented by an additional support and synthesis project. The second iteration of the programme, launched in 2011, has a similar approach, again with a set of training-focused projects supported by an additional synthesis effort.

The programme included this training strand in both iterations in order to deliver a holistic approach to improving research data management in the university context in which both human and technical infrastructures are addressed.

3. THE JISC MANAGING RESEARCH DATA PROGRAMME

3.1 The Training Strand: Aims and Vision

It is a principle of the JISC MRD programme that there is little benefit in building systems and technical infrastructure unless motivation, recognition and reward and data management skills among the research population are also addressed. For this reason it was felt necessary for projects to develop and embed RDM training materials in discipline-focused postgraduate courses to help make clear the benefits and rewards of effective research data management at an early stage in the research career.

3.2 UK Researcher RDM Skills Needs

The shortfall in data management training in UK higher education is widely recognised. A 2009 *Nature* editorial 'Data's shameful neglect' concluded that 'data management should be woven into

⁴ See <http://archive.stsci.edu/hst/bibliography/pubstat.html>. Observations by the Hubble Space Telescope are made on the basis of proposals, data is collected and made available to the proposers; data is stored at the Space Telescope Science Institute and made available after an embargo. Each year approx 200 proposals are selected from a field of 1,000; leading to c. 20,000 individual observations. There are now more research papers published on the bases of 'reuse' of the archived data than those based on the use described in the original proposal.

⁵ This is another area of significant activity in the JISC Managing Research Data Programme, but one which goes beyond the scope of the present paper.

every course, in science, as one of the foundations of knowledge' [8], a view which has found agreement elsewhere [9].

This acknowledged need to increase skills in managing research data among staff in HEIs, including researchers, librarians and research support staff, was explored by the UKOLN 'Dealing with Data' report of 2007 [10] and Swan and Brown's 2008 report on 'The skills, role and career structure of data scientists and curators'[11]. It was considered further in the second Research Data Management Forum of November 2008. These discussions were presented in the form of a white paper by Graham Pryor and Martin Donnelly, where the case is forcefully made that 'data skills should be made a core academic competency' and that 'data handling [should be] embedded in the curriculum'.[9]

Some UK organisations have attempted to address this shortfall. The Digital Curation Centre (DCC) has developed a wealth of digital curation and research data management training materials.⁶ The UK Data Archive provides extensive guidance and training materials on the creation, management and sharing of research data.⁷ Under its Researcher Development Initiative, the Economic and Social Research Council (ESRC) funded a 'Data Management and Sharing for Researchers Training Programme' which developed a programme of training for researchers and research support staff [12].

Additionally, under the heading 'Information Management', the Vitae Researcher Development Framework (Vitae RDF) includes the following description of necessary skills acquisition: 'Develops a sustained awareness of the creation, organisation, validation, sharing and curation of data.' [13] An 'Information Literacy Lens' [14] on the Vitae RDF, which includes considerable emphasis on data management skills, has been developed in consultation with the Research Information Network (RIN)'s Information Handling Working Group.⁸

Research presented in the RIN-funded report *To Share or Not to Share* [15] highlighted researchers' concerns and misgivings about making research data available for verification and reuse. Early findings from projects in the JISC Managing Research Data programme, moreover, highlighted awareness and skills gaps among researchers and called for advocacy, guidance and training materials to address these issues.⁹ Numerous reports have underlined the value of early intervention in the research career, including work by Sheila Corrall¹⁰, the JISC and others¹¹.

⁶ See <http://www.dcc.ac.uk/training>

⁷ See <http://www.data-archive.ac.uk/create-manage>

⁸ <http://www.rin.ac.uk/our-work/researcher-development-and-skills/information-handling-training-researchers/working-group-i> and see, e.g. <http://www.vitae.ac.uk/researchers/1271-414711/Learn-about-information-handling-lens-on-Researcher-Development-Framework.html>

⁹ See project outputs for JISC Incremental: <http://www.lib.cam.ac.uk/preservation/incremental/index.html>, JISC Sudamih: <http://sudamih.oucs.ox.ac.uk/>, JISC MaDAM: <http://www.library.manchester.ac.uk/aboutus/projects/madam/> and the University of Southampton: <http://www.southamptondata.org/>.

¹⁰ Sheila Corrall has recognised the importance of data literacy training at postgraduate student level in 'Roles and

Consonant with such initiatives and the concerns they reflect, it has been observed that there 'is a need to go beyond the workshop and the short training course, and embed preparation for a professional (and personal) lifetime of digital data curation within the academic curriculum'.¹²

3.3 Research Support Staff RDM Skills Needs

As well as integrating research data management skills in curricula for discipline specialists, it is also necessary to develop targeted course materials for librarians, research support staff and data managers. Calls for the 'upskilling' of subject or liaison librarians for roles which encompass support for the management and preservation of digital research data have become more urgent of recent years. In 2008, Alma Swan and Sheridan Brown observed that 'The role of the library in data-intensive research is important and a strategic repositioning of the library with respect to research support is now appropriate'. Swan and Brown envisaged three roles for the library with regard to research data management as a precondition to data intensive research. These were:

1. Increasing data awareness among researchers.
2. Providing archiving and preservation services.
3. Developing a new professional strand of practice in the form of data librarianship.[11]

Such analyses of the field highlight the importance of addressing the respective needs of researchers, librarians and research support staff. The importance of training for librarians and research support staff was clearly recognized when designing the first MRD programme in 2009-10, but it was judged that other agencies and stakeholders were able to take forward work to develop training materials and curricula to improve research data management among librarians, for example. It was felt that the initial priority should be to address the needs of postgraduate students and early career researchers as relatively little work had been done in those areas. While this prioritization may have been reasonable, with the benefit of hindsight it is acknowledged that an opportunity was missed to advance work to improve data management skills among librarians and other key research support staff at that point. Work in the second iteration of the Managing Research Data Programme is designed to address this shortfall.

3.4 The RDMTrain Projects

In the first iteration of the JISC MRD programme, the object of the five training projects, collectively known as 'RDMTrain', was to create materials which translated, where possible, generic training resources into something meaningful and targeted to postgraduate students studying in specific disciplines, and viewed as an essential part of training and research skills in these

responsibilities: Libraries, librarians and data'. In G Pryor (Ed.), *Managing research data* (pp. 105-133). London: Facet.

¹¹ For examples, see the Arcadia project report of work in this area at University of Cambridge: <http://arcadiaproject.lib.cam.ac.uk/docs/PINOTA-Report.pdf>, and the recommendations of the JISC/RIN/DCC DaMSSI final report, available at http://www.rin.ac.uk/system/files/attachments/JISCFinalreport_DaMSSI_FINAL.pdf.

¹² Pryor and Donnelly 2009, p.166.

disciplines. These materials were to be sustained by embedding in existing postgraduate training provision as well as being made openly available through the Jorum portal¹³.

The RDMTrain projects targeted the following disciplines: archaeology, the creative arts, geosciences, health sciences, psychology and social anthropology. A deliberate spread of disciplines was intentional: the programme did not intend to work only with scientific disciplines, which are often more familiar with discourse around the idea of data, but to also extend the terminology of data management into arts and humanities disciplines. The materials developed by the projects drew on user needs analysis undertaken with target audiences, and took the form of online guidance, practical software exercises, in-person training events and specimen data management plans alongside templates and supporting guidance materials.

3.5 The Data Management Plan

The RDMTrain projects of the first iteration of the MRD programme interrogated the DCC's Data Management Planning online tool [19] and its suitability for use within their target disciplines. They produced a set of discipline-focused templates for a data management plan (DMP), showing that discipline specificity, including the use of language appropriate to the discipline, encourages engagement with data management planning. However, further work is necessary to understand how data management planning can be optimised to the needs of a variety of disciplines and institutions.

The recent funding body mandates to embed data management planning as part of research practice can be useful to those providing training. Students wish to produce a DMP specifically relevant to them, often as a learning outcome of the course or as part of their wider skills development. Self-directed learning with access to customised guidance for the discipline and moderated exercises around the development of a DMP works well.

The DMP can be easily understood as another piece of administration which researchers are becoming obliged to complete. But the DMP can offer a research team a number of more sophisticated and engaging benefits when viewed as a dynamic tool which can be completed at the outset of the research work but regularly revisited during the work of the project to guide decision making about data use, re-use, storage and sharing. The DMP has potential as a pedagogical – or as one of the training projects suggested, andragogical – tool as, in order to be effective, data management planning must be an activity or learning process which draws on the experience of the working professional and informed by their experience in the role. Finding out the information required for the initial completion of the DMP helps the researcher to develop an awareness of the many issues connected to data management and leads to the ability for more sophisticated decision-making. This process can also provide a way to building the relationships between researchers and support staff which are required for the collaborative completion of the DMP; this can lead to new appreciation of the various roles involved in data management across the institution. In this way, the DMP also has the potential to influence researcher behaviour in regard to data management. In addition, the DMP is also a useful way of addressing the requirements of freedom of

information legislation, by providing evidence of an intention to release research data¹⁴.

The emphasis on data management planning is viewed by some funders and by the DCC as a core way of improving RDM practice. This seems a valid approach but there is still some work to be done on refining our understanding on what an optimal DMP – which aims to serve the requirements of a variety of stakeholders – might be.

3.6 DaMSSI Support and Synthesis

The five training projects of the first iteration were also accompanied by a support and synthesis project which was co-funded by the MRD programme and by the RIN, and was run with the co-operation of the DCC. This was the Data Management Skills Support Initiative ('DaMSSI') [16] which was overseen by the RIN Information Handling Working Group. One of DaMSSI's main purposes was to test the effectiveness of the Society of College, National and University Libraries (SCONUL)'s Seven Pillars of Information Literacy model [17] and the Vitae RDF for consistently describing data management skills and skills development paths in UK postgraduate courses. With the collaboration of the five projects, DaMSSI mapped individual course modules to the Seven Pillars and the Vitae RDF, and to the DCC digital curation lifecycle model [18] and identified basic generic data management skills alongside discipline-specific requirements. A synthesis of the training outputs of the projects was then carried out which investigated further the generic versus discipline-specific considerations and other successful approaches to training that had been identified as a result of the five projects' work.

In addition, DaMSSI produced a series of career profiles to help illustrate the fact that data management is an essential component - in obvious and less obvious ways - of a wide range of professions [16].

3.6.1 DaMSSI Findings and Recommendations

Finally, as a result of working with the RDMTrain projects, and in liaison with various wider stakeholders in data management and curation, DaMSSI formulated a set of recommendations for the institutions and projects embarking on future data management training development. These recommendations are based on synthesised feedback from the training strand projects about what factors contributed to the success of their training, and feedback received by the training projects from students whilst piloting their training offerings [19].

Some of the DaMSSI recommendations compared successful approaches in generic and discipline-specific approaches to data management training.

The first of these recommendations advised that those developing training work closely with disciplinary experts to ensure that terminology used within courses is accurate and clear to the target audience. This includes agreeing a basic definition of core concepts such as what 'data' can be within the discipline. This is particularly helpful for non-science disciplines.

¹³ <http://www.jorum.ac.uk>

¹⁴ See the guidance which specifies this requirement among others, from the Information Commissioner's Office, Sep 2011 - http://www.ico.gov.uk/news/latest_news/2011/ico-issues-advice-on-the-disclosure-of-research-information-26092011.aspx

Overviews and central descriptions of topic areas should be basic and generic, in order to introduce the topic at a level that is interesting but digestible for PhD students. This approach also allows modules to be more easily integrated into existing larger research methods courses.

In order to highlight relevance to the audience, however, generic material should be interlaced with discipline-specific examples, references and case studies wherever possible. This also helps to engage the audience, puts basic points into context and makes them understandable.

The RDMTrain projects found that training was more successful where training developers acknowledged accepted research practices within the discipline and worked to develop training materials that reflect these practices; for example, kinds of data handling, research funder expectations and popular archives and repositories.

Finally, training providers should use trainers with extensive knowledge of the discipline. Trainers who know the discipline well can provide the context and interlaced examples that engage students and make the topic seem relevant to them.

These observations raise important questions about training in research data management. Where, indeed, does such training ideally sit in the offering of a higher education institution, how is it most effectively delivered and who should be responsible for it? As a core research skill, intimately related to the practice of particular disciplines and affected by the specificities of the data collected, is it not right to argue that RDM should be tightly integrated with the postgraduate (or even undergraduate) training of a given discipline? Here for example, we might allude to the training in excavation and recording practice received by archaeologists, the knowledge of survey design and statistical analysis necessary among social scientists and the requirements among chemists and other experimental scientists to maintain a lab notebook. Is not RDM simply a core part of good research practice, which, along with other skills, should be inculcated early in the core disciplinary training of research students?

However, another point of view might be that RDM is a generic skillset, applicable to all disciplines. If RDM is regarded as a branch of information literacy, might it not be more effective and efficient to offer training alongside other such skills that are often delivered centrally, by staff that are specialists in approaches to information management? Recent studies [20, 21] of information handling among postgraduate students seem to suggest that there is a genuine, if not to say urgent, need for specific training in information handling skills and this cannot reliably be left to discipline specialists.

These considerations are fundamental and not susceptible to immediate solutions, particularly as we are at an early stage of integrating RDM training in curricula. Many universities will have to dose any solution with a generous helping of pragmatism. The JISC RDMTrain projects, DaMSSI, the RIN-led coalition and other stakeholders believe it is vitally important to promote RDM training and to share practice around delivery as this develops.

Another key group of the DaMSSI recommendations address the issues around language used in researcher training for data management. As identified in earlier JISC MRD programme

work¹⁵, the language and terminology used in the presentation of guidance and of training can make a significant difference in the extent to which researchers see the material as relevant to their discipline and engage with support infrastructure to better manage their data. The DaMSSI project found that, ‘echoing the findings of the earlier JISC MRD Incremental project, many researchers don’t understand much of the specialist language from the information or preservation worlds’ [19]. These issues continue to be explored in the work of the JISC-funded SHARD project.¹⁶

Language issues arose again when DaMSSI worked with the training projects to ascertain use cases for the SCONUL Seven Pillars and Vitae Researcher Development Framework models. In the first instance, many project staff members were confused by the acronym ‘RDF’ for the Researcher Development Framework, this acronym already being widely understood in this community to denote a completely different concept. In addition, each of the Seven Pillars has a name that has immediate relevance to data management, but the definition of these terms is at times different for different audiences. For example, the ‘Plan’ pillar in the Seven Pillars model focuses specifically on search strategies for locating information, whilst ‘plan’ within a data management lifecycle has a broader and earlier definition of planning how data will be managed at the same time as a research project is outlined. That process, however, would currently be more aligned within the Seven Pillars model with the ‘Scope’ pillar.

DaMSSI recommended that training providers should avoid using acronyms and data curation-specific terminology, and instead explain principles and issues in language that is understandable to a general audience and is not already weighted for the audience’s discipline: for example, the term ‘curation’ already has specific meaning for much of the creative arts.

It is hoped that these recommendations will contribute to the subsequent development of successful postgraduate-level RDM training materials.

4. FUTURE ACTIVITY

Activities in the second JISC Managing Research Data Programme to address training requirements are driven by the findings of the first programme and the recommendations of the DaMSSI project. There has also been an effort to cover areas relatively neglected in the first programme and to respond to changing circumstances.

4.1 Future RDM Responsibilities: Cross-Campus

The EPSRC’s ‘Policy Framework’ and ‘Expectations’ [23] for RDM have underlined what was already a growing recognition that solutions to the research data challenge will require ‘cross-campus’ responses which coordinate a number of stakeholders, including researchers, the library, computing services and research support services. Although much responsibility for research data management must necessarily remain with the individual researcher, PI or research group, it has been recognized that various agencies within universities and research institutions

¹⁵ Namely the JISC Incremental project at the Universities of Cambridge and Glasgow, project website available at: <http://www.lib.cam.ac.uk/preservation/incremental/index.html>.

¹⁶ This project blogs at <http://shard-jisc.blogspot.co.uk/>

have important supporting roles to play. This realisation coincides with increasingly urgent calls for university libraries to adapt to the requirements of research as it becomes more data-centric. The recent report by Mary Auckland for Research Libraries UK, appropriately entitled *Re-Skilling for Research*, communicates a palpable sense of urgency:

A shift can be seen which takes Subject Librarians into a world beyond information discovery and management, collection development and information literacy training, to one in which they play a much greater part in the research process and in particular in the management, curation and preservation of research data, and in scholarly communication and the effective dissemination of research outputs. [24]

The Dutch 3TU Datacentrum, a collaborative effort between the Netherlands' three technical universities, has developed the 'Data Intelligence 4 Librarians' course for which there is a substantial amount of online material [25]. The programme aims to equip librarians better to 'to advise researchers effectively and efficiently' in data curation. Such work is extremely useful, but there remains – as in the case of researchers themselves – a need to embed training in research data management skills in Library and Information Science postgraduate courses in order to ensure such skills are a *sine qua non* for the next generation of librarians. With these issues in mind, the Managing Research Data programme has, in its second iteration, explicitly targeted the development of training materials for librarians, funding a project led by the University of Sheffield iSchool.

4.2 RDMTrain 02

By and large, the training projects in the first Managing Research Data programme focused on the arts, humanities and social sciences. This orientation stemmed from a number of related considerations: the opportunity to build on existing materials coincided with a tangible need for skills development and an estimation that the challenges in these subject areas, while significant, may yet be relatively tractable. There has also been feeling that the more technical focus of STEM subjects – and the higher levels of funding available – meant that JISC-funded work was less necessary and would have a less tangible impact. However, reports such as the RIN study into *Information Practices in the Physical Sciences* [26] suggest that such assumptions may, at least in part, be misplaced. The second iteration called for projects to develop materials in subject areas which had not been covered in the first programme, and it is notable that projects targeting more technical subjects were prominent among those funded and include computer science, digital music research, physics and astronomy.

4.3 DaMSSI-ABC

As a whole, and specifically through the new DaMSSI-ABC support project, the training strand of the JISC Managing Research Data programme seeks to promote the incorporation of RDM components into the training of current and future researchers and research support staff. Building on the findings and recommendations of the first programme, the second iteration seeks in particular to ensure that materials are as reusable as possible and to promote them with learned societies and professional bodies.

'ABC' in the support project's name stands for Assessment, Benchmarking and Classification, underlining a commitment to

ensuring that the training materials developed are as discoverable, reusable and interoperable as possible. With the assistance of the support project, the programme will work closely with the JISC-funded Jorum repository for reusable learning and teaching materials (or in Jorum terminology, Open Educational Resources). In collaboration with the MRD programme, Jorum will be piloting a research data management-flavoured portal in order to assist access to training materials [27]. The motivation behind this activity is to a) draw attention to research data management as an important component of more general research skills, and b) make the related materials more easily discoverable and reusable.

An essential component of reusability, when it comes to learning and teaching resources including training materials, is understanding precisely how the material might be incorporated into existing courses of diverse characteristics. Standardised descriptions, mapping of assessment, benchmarking required attainments and detailing subsequent classification are arguably the necessary components for the interoperability of training materials. A central focus of the DaMSSI-ABC project will be to make practical and informative recommendations on the basis of examining UK and international frameworks for benchmarking and classifying training materials. Existing models include the US Library of Congress's Digital Preservation Outreach and Education (DPOE) audience classification pyramid¹⁷ (which may provide a useful guide for identifying courses aimed at executive-level strategic planners, operational managers and practitioners) and the Vitae RDF, but other initiatives will be taken into account, as well as the expertise of key UK stakeholders.

4.3.1 DaMSSI-ABC: The Role of Learned Societies

The important role of learned societies and professional bodies in contributing to the formulation of training materials, endorsing them and promoting them as part of the development support that they offer to their members is clearly recognised. As custodians of professional standards, these bodies are obvious interlocutors for the purpose of helping to promote data management skills, and to get these skills better recognised by students and researchers as indispensable elements in their career development. However, most such bodies have had little or no involvement in information literacy issues. The DaMSSI-ABC project in its support role will work to encourage and facilitate a dialogue between the funded projects and appropriate learned societies / professional bodies. This work will aim to ensure that data management skills are recognized by relevant learned societies and professional bodies as an indispensable part of researchers' career development, to accordingly identify data management champions within these organizations and to involve them in identifying means of skills assessment and benchmarks.

4.3.2 DaMSSI-ABC: Other Planned Activity

Other principle areas of activity include:

- Encouraging the early encounter with research data management issues in the research career;
- Working to help researchers and research support staff to plan career development;

¹⁷ A useful description and diagram of the DPOE pyramid, along with definitions of each audience, is available at <http://www.digitalpreservation.gov/education/educationneeds.html>

- Exploring ways to assess and compare courses; and,
- Reporting, where possible, on diverse strategies for incorporating RDM training in discipline-specific curricula or more generic research skills offerings.

In this way, the DaMSSI-ABC project aims to contribute to the uptake and reuse of RDM training materials in the UK (and potentially internationally) as well as increasing our understanding of the most effective description, benchmarking and classification of such materials.

5. CONCLUSION

This paper relates the efforts of the JISC Managing Research Data programme to encourage the development and uptake of RDM training materials across UK institutions and disciplines. Although much progress has been made, the authors are obliged to recognize that considerable work is still required before research data management training is widely incorporated into postgraduate training curricula.

It is hoped that this paper will contribute to an international debate around the place of research data management training, and how it may best be delivered and promoted. We particularly emphasise the value of emerging work a) to engage learned societies and professional bodies; b) to establish practical and effective means of describing, benchmarking and classifying training materials to promote reuse; and c) to encourage colleagues across the university campus to engage with research data management and to tackle its challenges in collaboration.

6. ACKNOWLEDGMENTS

Our thanks to all the project staff on the CAIRO, DataTrain, DATUM for Health, DMTpsych and MANTRA projects, the RIN Information Handling Working Group (now the Research Information and Digital Literacies Coalition), and our colleague Kellie Snow of HATII at the University of Glasgow, for their essential contributions to the findings of this paper.

7. REFERENCES

- [1] Research Councils UK (2011). *RCUK Common Principles on Data Policy*. Available at: <http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx>
- [2] Organisation for Economic Co-operation and Development (2007). *OECD Principles and Guidelines for Access to Research Data from Public Funding*. Available at: <http://www.oecd.org/dataoecd/9/61/38500813.pdf>
- [3] European Federation of National Academies of Sciences and Humanities (2012), ‘Joint declaration as European Commission VP Neelie Kroes and ALEA agree on Open Science, Rome, 11-12 April 2012’. Press release, published 11th April 2012. Available at: <http://www.allea.org/Pages/ALL/33/144.bGFuZz1FTkc.html>
- [4] The Royal Society (2012). *Science as an Open Enterprise: Final Report*. Available at: <http://royalsociety.org/policy/projects/science-public-enterprise/>
- [5] EPSRC (2011). *Policy Framework on Research Data: Principles*. Available at: <http://www.epsrc.ac.uk/about/standards/researchdata/Pages/principles.aspx>
- [6] Dryad (2011, updated 2012). *Joint Data Archiving Policy*. Available at: <http://datadryad.org/jdap>
- [7] Opportunities for Data Exchange (2011). *Report on Integration of Data and Publications*. Available at: http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2011/11/ODE-ReportOnIntegrationOfDataAndPublications-1_1.pdf.
- [8] *Nature* (2009). ‘Editorial: Data’s shameful neglect’. 461, p. 145. Published online 9 September, 2009. Available at: <http://www.nature.com/nature/journal/v461/n7261/full/461145a.html>. doi:10.1038/461145a
- [9] Pryor G and Donnelly M (2009). ‘Skilling Up to Do Data: Whose Role, Whose Responsibility, Whose Career?’ *International Journal of Digital Curation*, Vol. 4, No. 2, pp. 158-170. doi:10.2218/ijdc.v4i2.105
- [10] Lyon L (2007). *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, UKOLN consultancy report. Available at: http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.doc
- [11] Swan A and Brown S (2008). *The Skills, Role and Career Structure of Data Scientists and Curators: an Assessment of Current Practice and Future Needs: Report to the JISC*. Available at: <http://www.jisc.ac.uk/media/documents/programmes/digitalRepositories/dataskillsCareersFinalReport.pdf>
- [12] Lyon L, Van den Eynden V and Bishop E (2011). ESRC Researcher Development Initiative, ‘Data Management and Sharing for Researchers Training Programme’. Outputs available at: <http://www.esrc.ac.uk/my-esrc/grants/RES-046-25-0038/read>
- [13] Vitae (2011). *Researcher Development Framework*. Available at: <http://www.vitae.ac.uk/policy-practice/165001/Researcher-development-framework-consultation.html>
- [14] Vitae (2012). *Information Literacy Lens on the Researcher Development Framework*. Available at: http://www.vitae.ac.uk/CMS/files/upload/Vitae_Information_Literacy_Lens_on_the_RDF_Apr_2012.pdf
- [15] Swan A and Brown S (2008). *To Share or Not to Share: Publication and Quality Assurance of Research Data Outputs. A report commissioned by the Research Information Network (RIN)*. Available at: <http://rinarchive.jisc-collections.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs>
- [16] Research Information Network (2010). ‘Data Management, Information Literacy and DaMSSI’. Available at: <http://rinarchive.jisc-collections.ac.uk/our-work/researcher-development-and-skills/data-management-and-information-literacy>
- [17] SCONUL (2011). *The SCONUL Seven Pillars of Information Literacy Core Model for Higher Education*. Available at: http://www.sconul.ac.uk/groups/information_literacy/publications/coremodel.pdf

- [18] Digital Curation Centre (2010). *The Digital Curation Lifecycle Model*. Available at: <http://www.dcc.ac.uk/resources/curation-lifecycle-model>
- [19] Molloy L and Snow K (2011). *DaMSSI Final Report*. Available at: http://www.dcc.ac.uk/webfm_send/532 and http://rinarchive.jisc-collections.ac.uk/system/files/attachments/JISCfinalreport_DaMSSI_FINAL.pdf
- [20] Research Information Network (2009). *Mind the Skills Gap: Information-Handling Training for Researchers*. Available at: <http://rinarchive.jisc-collections.ac.uk/our-work/researcher-development-and-skills/mind-skills-gap-information-handling-training-researchers>
- [21] JISC and the British Library (2012). *Researchers of Tomorrow: The Research Behaviour of Generation Y Doctoral Students*. Available at: <http://www.jisc.ac.uk/publications/reports/2012/researchers-of-tomorrow.aspx>
- [22] Digital Curation Centre (2007). *Data Management Online Planning Tool*. Available at: <https://dmponline.dcc.ac.uk/>
- [23] Engineering and Physical Sciences Research Council (2011). *EPSRC Expectations*. Available at: <http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx>
- [24] Auckland, M (2012) for Research Libraries UK. *Re-skilling for Research*. Available at: <http://www.rluk.ac.uk/content/re-skilling-research>
- [25] 3TU.Datacentrum (2012). *Data Intelligence 4 Librarians* course available at: <http://dataintelligence.3tu.nl/en/news-events/news-item/artikel/data-intelligence-4-librarians-online/>
- [26] Research Information Network (2011). *Collaborative yet Independent: Information Practices in the Physical Sciences*, report for RIN and the Institute of Physics. Available at: <http://www.rin.ac.uk/our-work/using-and-accessing-information-resources/physical-sciences-case-studies-use-and-discovery->
- [27] Siminson N (2012). ‘Piloting institutional and subject flavours of Jorum’. Blogpost, published online 19 April 2012. Available at: <http://jorum.ac.uk/blog/post/30/piloting-institutional-and-subject-flavours-of-jorum>

Ahead of the CurV: Digital Curator Vocational Education

Laura Molloy

Humanities Advanced Technology and Information
Institute (HATII)
University of Glasgow
Glasgow, Scotland
(+44) (0)141 330 7133

Laura.Molloy@glasgow.ac.uk

Ann Gow

Humanities Advanced Technology and Information
Institute (HATII)
University of Glasgow
Glasgow, Scotland
(+44) (0)141 330 5997

Ann.Gow@glasgow.ac.uk

ABSTRACT

In this paper, we describe the work of the EC-funded DigCurV project. We examine the context of the project, the methods and findings of its extensive survey work, and the development of proposed frameworks for evaluating and delivering a digital curation curriculum.

Keywords

Training, education, skills, skills frameworks, vocational training, lifelong learning, curriculum development, digital curation, Europe.

1. INTRODUCTION

The Digital Curator Vocational Education (DigCurV) project [1] is funded by the European Commission (EC)'s Leonardo da Vinci lifelong learning programme [2] and will run until the end of June 2013. It aims to establish a curriculum framework for vocational training in digital curation.

DigCurV brings together a network of partners to address the availability of vocational training for digital curators in the library, archive, museum and cultural heritage sectors, with a particular focus on the training needed to develop new skills that are essential for the long-term management of digital collections.

2. BACKGROUND

A current and increasingly urgent issue within the cultural heritage sector across Europe and north America is the ability of conservators, researchers and other staff to effectively care for digital objects now appearing within their collections. But are those now professionally obliged to perform digital curation receiving the training they need? And what exactly constitutes those training needs?

Another dilemma arises when attempting to define who is responsible for the long term preservation and curation of digital objects held by an institution. The Italian economist Vilfredo Pareto argued at the turn of the twentieth century that a society

grown wealthy enough would cease to foster general knowledge in individuals and increasingly encourage individual ability in tightly specified and increasingly complex skills. Each worker would become increasingly proficient at one element of the work of a larger project or process. We are currently at a similar point of decision with digital curation training. Should all workers in the cultural heritage sector become more proficient in the curation of digital assets for which they are responsible, or should we be planning intensely specific training to enable a distinct strain of specialists to emerge? It is in the context of these debates that DigCurV is working.

DigCurV is using the thirty months of the project, which began in January 2011, to identify, analyse and profile existing training opportunities and methodologies, survey training needs in the sector and identify the key skills and competences required of digital curators. The project then aims to establish a curriculum framework from which training programmes can be developed. This curriculum framework will be tested and then published in at least four languages. A final conference for policy and decision makers is planned to raise awareness of the curriculum and promote it to those developing training, evaluating training and planning to undertake training.

3. AN INTERNATIONAL NETWORK

The DigCurV project brings together a network of partners from across Europe and north America to capitalise on expertise and experience in training across national and linguistic boundaries. Founding project partners come from Italy, Scotland, Ireland, Germany, Lithuania, England, Canada and the USA.

HATII at the University of Glasgow is a key partner in the DigCurV effort. HATII is a multidisciplinary research and teaching institute in the fields of digital curation, digital humanities, archives, records and information management and other areas connected to the use and management of digital information. Members of the research team at HATII have been central to the work of various other UK and European research projects in digital curation, digital preservation and research data management [3], and also contribute teaching, curriculum development and training development experience to DigCurV.

In addition to the network of founding partners, the DigCurV community includes an extensive network of members worldwide, including forty-four cultural heritage institutions and eighty-six individuals.

4. THE NEED FOR VOCATIONAL TRAINING

The EC has encouraged the growth of digital information professions with the 2005 launch of its i2010 strategy and a subsequent Digital Agenda initiative, launched in 2010 [4].

This investment is justified by the importance of the cultural heritage sector in the European economy. Specifically, in addition to the thousands of universities, libraries and archives across Europe, there are also more than 19,000 museums and art galleries, which employ around 100,000 staff [5]. Traditionally, museums and gallery staff have been trained in physical object care by well-established professional and vocational training courses, but as digital technologies infiltrate every aspect of society, digital objects are increasingly making their way into the collections held by memory institutions.

In 2004, the Digital Preservation Coalition (DPC) and the UK Joint Information Systems Committee (JISC) established the need for digital preservation skills training in multiple sectors in the UK [6], and DigitalPreservationEurope (DPE) research has also echoed the need for these skills to be regularly refreshed by professionals as digital curation practice develops and evolves [7]. In 2009, the New York Times recognised the growing demand for digital archivist skills in the USA [8]. In 2010, Gartner Research identified four new roles needed by IT departments to remain effective [9] – one of these was ‘digital archivist’, and it was then estimated that fifteen percent of businesses would employ in this role by 2012. And yet, at the 2011 JISC International Curation Education (ICE) Forum in the UK [10], fewer than half a dozen UK institutions were listed as providing digital curation training as part of their profession library and archive courses.

However, it is not enough to trust new recruitment into the cultural heritage sector to face the challenges of digital curation. Research conducted by DigCurV confirms that at least in the experience of our respondents, investment is not always channelled towards creating new staff to take on the emerging digital curation duties increasingly required by heritage institutions. There is a need for existing staff in cultural heritage institutions to adapt to the emerging digital cultural sector.

5. TRAINING NEEDS, OPPORTUNITIES AND SKILLS

DigCurV started work by consulting research already undertaken in the area of digital curation training, in order to move toward the development of a useful and usable curriculum framework for new and existing digital curation professionals in the cultural heritage sector. Data was then gathered about current training needs and training opportunities. An online registry of training was established on the project website to promote available training to the public [11]. The project organised focus groups in several European countries to talk to professionals in the field about the skills currently required by various contemporary roles in digital curation, and performed analysis of a number of job advertisements to establish currently required skills. An evaluation framework was developed to apply to existing training. These activities all influence the development, currently underway, of the initial curriculum for digital curation vocational training.

5.1 Training Needs Survey

The training needs survey ran online during July and August 2011, and received 454 valid responses from 44 countries, mostly from Europe. Most responding institutions are currently undertaking long-term preservation of digital materials. More than half, however, reported they were not aware of recruitment of new staff to undertake these responsibilities, thereby implying that existing staff will be obliged to acquire the necessary skills and competences.

A significant proportion of respondents reported their organisation is currently planning training for staff in digital curation skills, with small group workshops the most popular format.

The survey team identified skills for digital curation from scrutiny of the literature and previous research on this topic, including reference to the OAIS Reference Model [12], the Digital Curation Centre (DCC) Lifecycle Model [13], the Digital Preservation Outreach and Education (DPOE) training needs assessment survey [14], Scheffel, Osswald and Neuroth’s work on qualification in digital preservation [15] and work by Kim, Addom and Stanton on education for e-science [16].

From the resulting long list of general skills, almost all were regarded by respondents as relevant to digital curation work, but the ability to collaborate with others, the ability to communicate with others and an affinity for working with technology emerged from survey responses as the most prized in digital curation staff. A list of skills specific to digital curation work was also provided; again, virtually all were indicated as of high importance by respondents, but the highest need for training appeared to be in providing staff with a basic knowledge of digital curation issues, planning for digital preservation and data management, and the use of specific tools for digital curation.

These results chime with the view reached at the JISC ICE forum in 2011 that working competently in digital curation requires a high degree of competence in a markedly diverse set of skills.

5.2 Training Opportunities Survey

The training opportunities survey was distributed from April to June 2011. The main objectives of the survey were to identify, analyse and profile existing training opportunities. The survey included basic questions about the responding institution, but focused on issues related to training content, methodologies, delivery options, and assessment, certification and best practices for training and continuous professional development. Sixty valid responses were received from sixteen countries, again mostly from Europe.

Forty percent of respondents reported having arranged digital curation training events in the two years prior to the survey.

Most events were offered in western Europe and the US, and predominantly in capital cities with the exceptions of Germany and the UK. Almost half of all reported courses were delivered in English, although we are aware that the fact the training opportunities survey was conducted in English may have influenced this figure.

The most frequently-trained audiences were practitioners and researchers from archives, libraries, museums or academic institutions. Forty-eight percent of all training was appropriate for developers employed by commercial vendors or institutional IT experts within the museums, libraries, archives, government and

business sectors, who are responsible for digital curation. Thirty-three percent of reported training events were targeted at students from various sectors. Fifty-seven percent of reported courses required some basic understanding of the main principles of digital curation beforehand.

Skills were addressed again in this survey. Knowledge of key needs and challenges was seen as most important, followed by standards and strategic planning. Technical issues were taught in almost half of courses, followed by legal aspects, digital curation and preservation tools, digital repository audit and certification, and trusted repositories.

The training opportunities survey revealed gaps in training provision, particularly in eastern Europe and the Nordic countries. There may also be a lack of training in languages other than English. Survey responses emphasised how much existing training is focused on basic and introductory principles, with much less available for experienced practitioners in digital curation and digital preservation.

5.3 Skills Analysis: Focus Groups

In addition to the active programme of survey work, five project partner countries hosted focus groups in the period from September to November 2011. These groups aimed to identify the skills and competences needed for digital curation, what the relevant professional roles in digital curation were, and the corresponding training needs. Working with the DPOE audience pyramid [17], participants identified as practitioners, managers or executive, and presented a fairly consistent set of findings across countries and staff types.

Participants reported a lack of appropriately-skilled staff, presenting challenges to successful and timely recruitment. The diversity of the required skill-set echoed the survey findings; the ideal skill-set identified by participants combines technical expertise, information science, library or archival knowledge and subject knowledge along with strong communication skills.

Participants also reported a lack of suitable training which needs to be addressed with some urgency. The Irish and Lithuanian groups particularly reported the need for training in the introductory concepts of digital curation and preservation.

Participants were asked their opinion on the need for accreditation of training. Many were in favour: for practitioners as proof of their qualification, and for managers and executives as a benchmark useful during the recruitment process. Other participants from the manager staff group, however, held the opinion that skilled staff are so urgently needed, they would prioritise possession of the relevant skills above accredited status during the recruitment process, and so there was no decisive view across those interviewed.

5.4 Skills Analysis: Job Advertisements

Forty-eight job advertisements, representing fifty-three openings, were collected in the period from February 2011 to January 2012, from the UK, USA, New Zealand, Germany and Australia. This exercise was to provide a snapshot of the current state of recruitment in the digital curation sector, as opposed to any attempt at a representative collection.

These were scrutinised for the skills required; competences, experience and knowledge expected; and the degrees and qualifications asked for. The tasks expected by the incumbent were also noted. The findings of this activity again echo the

messages emerging from the other research undertaken by the project team.

Classifying skills into ‘general’ and ‘digital curation-specific’, as with the training needs survey, the team found that once again the role demands an extensive set of diverse abilities.

The most frequently cited ‘general’ tasks listed as essential to the role included communications including outreach and liaison, project management, teaching and training, supervision and funding capture. The most popular digital curation-specific tasks were digital collection management, data management, broad-based digital curation and preservation, trusted repository and archive-appropriate duties, documentation of assets and awareness of best practice.

The skills, competences and knowledge sought from applicants were again considered in two separate groups by the research team. The most commonly cited ‘general’ skills were communication, collaboration and team work. Popular digital curation-specific skills included knowledge of digital archive and digital library environments, trusted repositories, lifecycle data management, information technology in general, programming, metadata, up-to-date experience of digital preservation tools and policies, awareness of current standards and knowledge of best practice.

An advanced degree, usually master’s degree or equivalent, was the most desirable qualification, and preferably from library and information studies or archive courses, a science discipline, computer science or humanities.

6. EVALUATION FRAMEWORK

Such extensive research was a salient element of the approach to the development of an evaluation framework. The findings of our research described the current skills and training landscape, including which skills were most sought by those in the profession, the availability of individuals possessing these skills, and the current access to useful training for both new and existing staff. Many members of the DigCurV team have prior experience in digital curation, data management and skills training work, and so could contribute experience of UK and international projects and initiatives such as DPE, nestor, Planets, the JISC Managing Research Data programme and its Data Management Skills Support Initiative (DaMSSI) and the DCC, amongst others. This massed experience further informed our view of the current landscape, providing us with a profile of digital curation training, which we further augmented by drawing on the findings of other work that has already taken place in digital curation skills training.

On the basis of these sources of information, we developed an evaluation framework, which is intended to be helpful to those providing or assessing digital curation curricula (or individual pieces of training which may form part of a curriculum). The layout is based on the matrix format of the DigCCurr Matrix of Digital Curation Competencies and Knowledge [18]. Other models drawn upon include the DPOE training audiences pyramid, the Digital Curation Centre lifecycle model and the Information Literacy Lens [19] developed by the Research Information Network (RIN) to apply to the Vitae Researcher Development Framework [20] and the Society of College, National and University Libraries’ Seven Pillars of Information Literacy model [21].

The DigCurV Evaluation Framework provides a series of different ways to view and evaluate a digital curation curriculum or piece of training. Taking a structured approach to consideration of a curriculum or piece of training can help to assess what training is already available, and to clarify which potential approaches, audiences and skills may need to be addressed. For those assessing training, the Evaluation Framework aims to provide a structure to which training offerings can be mapped. This serves to clarify where provision is ample and which approaches, audiences or skills are scarcely served in existing training. Mapping can also provide a benchmark to allow comparison of different training offerings against each other.

The Evaluation Framework prepares the ground for the subsequent Curriculum Framework, emerging later in the DigCurV project, which – as the name suggests – moves on from evaluating and reviewing existing training to assisting in the development of new training offerings.

7. DRAFT CURRICULUM FRAMEWORK

The DigCurV Curriculum Framework aims to indicate core digital curation skills and competences, and pathways of skills progression through these. It is not an attempt to specify a particular training curriculum, but instead is deliberately a reference framework. The Curriculum Framework will take the form of a portfolio document, comprised of three ‘lenses’ or views, one for each of the DPOE audience types: Practitioner (Lens 1); Manager (Lens 2) and Executive (Lens 3).

In each lens, skills and competences specified are based on the findings of the RIN Researcher Development Framework Information Literacy Lens, and populated with results of both the DigCurV training needs survey and DigCurV focus group findings. Within the Skills and Competences Matrix, the ‘Skills Area’ and ‘Descriptor’ columns are drawn from those in the RIN Lens which are applicable to digital curation. We are considering how practical, managerial and executive roles in digital curation map to each Descriptor. These skills and competences encompass not just technical knowledge and duties but widen out to also encompass personal attributes, attitudes and behaviours, further helping to define the approaches that a curriculum should encourage in individuals to shape them for success in digital curation professions.

Each lens aims to answer the question, ‘When building digital curation training for this level of staff in a cultural heritage institution, what should be included?’

The development of each lens draws on the consolidated experience and knowledge of the DigCurV team across all partners. Led by HATII, this work particularly relies on the teaching experience of the team as well as awareness of ongoing modelling of the RIN Information Literacy Lens promoted by HATII participation in the RIN Information Handling Working Group (now the Research Information and Digital Literacies Coalition).

8. FUTURE DEVELOPMENT

The answer to the dilemma of whether all cultural heritage professionals should up-skill in digital curation, or whether it should be left to specialists, is not the responsibility of one project such as DigCurV. Pragmatically, then, in order to address as many futures in digital curation as possible, the project continues to work with an open definition of lifelong learning and

vocational training, acknowledging the relevance of all postgraduate and professional-level training available to those already working in the field. This includes training types from short courses on specific skills for existing professionals in the sector, to master’s courses specifically training students in digital curation skills.

The international network established by the project – which includes and extends beyond the founding partners – will be involved in iterative development of the Curriculum Framework and will be specifically asked to participate in evaluation events in the second half of 2012. In addition, the project plans to consider domain-specific curricula, extend community use – both as contributors and browsers – of the training registry, and consider the feasibility of accreditation of training offerings.

9. ACKNOWLEDGMENTS

Our thanks to the DigCurV project team, and all our research participants.

10. REFERENCES

- [1] <http://www.digcur-education.org/>
- [2] European Commission Leonardo da Vinci programme: http://ec.europa.eu/education/lifelong-learning-programme/ldv_en.htm.
- [3] <http://www.gla.ac.uk/hatii>
- [4] European Commission’s Europe’s Information Society webpage available at: http://ec.europa.eu/information_society/eeurope/i2010/index_en.htm.
- [5] European Group on Museum Statistics estimate. Data is available at http://www.egmus.eu/index.php?id=88&no_cache=1.
- [6] JISC and DPC Training Needs Analysis: <http://www.jisc.ac.uk/media/documents/programmes/preservation/trainingneedsfinalreport.pdf>.
- [7] Harvey, R. (2007). *Professional Development in Digital Preservation: a life-long requirement*, DPE briefing paper
- [8] De Aenlle, C. ‘Digital Archivists: Now in Demand’, *New York Times*, 7th February 2009. Available at <http://www.nytimes.com/2009/02/08/jobs/08starts.html>.
- [9] Gartner Identifies Four Information Management Roles IT Departments Need to Remain Effective, press release available at <http://www.gartner.com/it/page.jsp?id=1282513>.
- [10] http://www.jisc.ac.uk/whatwedo/programmes/preservation/ic_eforum
- [11] <http://www.digcur-education.org/eng/Training-opportunities>
- [12] <http://public.ccsds.org/publications/archive/650x0b1.pdf>
- [13] <http://www.dcc.ac.uk/resources/curation-lifecycle-model>
- [14] <http://www.digitalpreservation.gov/education/documents/DPOENeedsAssessmentSurveyExecutiveSummary.pdf>
- [15] Scheffel, Osswald and Neuroth (no date), ‘Qualifizierung im Themenbereich „Langzeitarchivierung digitaler Objekte“ in Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. Goettingen, Germany. Available at http://nestor.sub.uni-goettingen.de/objekte/qualifizierung_im_themenbereich_langzeitarchivierung_digitaler_objekte_in_eine_kleine_enzyklopedie_der_digitalen_langzeitarchivierung.pdf

- goettingen.de/handbuch/artikel/nestor_handbuch_artikel_46_8.pdf
- [16] Kim, Addom and Stanton (2011). ‘Education for E-Science Professionals: Integrating Data Curation and Cyberinfrastructure’ in *International Journal of Digital Curation*, 1, 6: 125..Available at
<http://www.ijdc.net/index.php/ijdc/article/view/168/236>
- [17] <http://www.digitalpreservation.gov/education/educationneeds.html>
- [18] Version 13, 17 June 2009. Available at
<http://ils.unc.edu/digccurr/products.html>
- [19] <http://www.vitae.ac.uk/policy-practice/375-533141/The-Informed-Researcher-Booklet-and-Information-literacy-lens-on-the-Vitae-Researcher-Development-Framework-now.html>
- [20] <http://www.vitae.ac.uk/researchers/430901-291181/Researcher-Development-Framework-RDF.html>
- [21] https://www.sconul.ac.uk/groups/information_literacy/seven_pillars.html

Preserving Electronic Theses and Dissertations: Findings of the *Lifecycle Management for ETDs* Project

Martin Halbert
University of North Texas
1155 Union Circle #305190
Denton, TX, 76203
940-565-3025
martin.halbert@unt.edu

Katherine Skinner
Educopia Institute
1230 Peachtree Street
Atlanta, GA 30309
404-783-2534
katherine@metaarchive.org

Matt Schultz
MetaArchive Cooperative
1230 Peachtree Street
Atlanta, GA 30309
616-566-3204
matt.schultz@metaarchive.org

ABSTRACT

This paper conveys findings from four years of research conducted by the MetaArchive Cooperative, the Networked Digital Library of Theses and Dissertations (NDLTD), and the University of North Texas to investigate and document how academic institutions may best ensure that the electronic theses and dissertations they acquire from students today will be available to future researchers..

Categories and Subject Descriptors

E.1 [Data Structures]: *distributed data structures*. H.3.2 [Digital Libraries]: *Information Storage, file organization*. H.3.4 [Systems and Software]: *distributed systems*. H.3.6 [Library Automation]: *large text archives*. H.3.7 [Digital Libraries]: *collection, dissemination, standards, systems issues*.

General Terms

Management, Documentation, Performance, Design, Reliability, Standardization, Languages, Theory, Legal Aspects, Verification.

Keywords

Archival Information Packages, Data Management, Digital Archives, Digital Curation, Digital Libraries, Electronic Theses and Dissertations, ETDs, Digital Objects, Digital Preservation, Distributed Digital Preservation, Ingest, Interoperability, Micro-Services, Repository Software, Submission Information Packages.

1. INTRODUCTION

One of the most important emerging responsibilities for academic libraries is curatorial responsibility for electronic theses and dissertations (ETDs) which serve as the final research products created by new scholars to demonstrate their scholarly competence. These are important intellectual assets both to colleges and universities and their graduates. Because virtually all theses and dissertations are now created as digital products with new preservation and access characteristics, a movement toward ETD curation programs in both U.S. institutions and abroad began in the early 1990's and has continued to this day.

There are many articles documenting this movement. The Coalition for Networked Information (CNI) recently studied the history of ETDs and graduate education and conducted an international survey concerning ETDs that examined linkages between the growth of ETD programs, institutional repositories, open access and other important trends in higher education (Lippincott and Lynch, 2010). Additional key issues identified in

the CNI survey are questions and uncertainty within institutions concerning ETD embargoes, ETD format considerations, costs of ETD programs, and the role of libraries in working with graduate schools to maximize benefits of ETD programs for students.

A basic point made by the CNI study and virtually all current literature on the ETD movement is that colleges and universities have been steadily transitioning from traditional paper/microfilm to digital submission, dissemination, and preservation processes. Increasingly, academic institutions worldwide are now accepting and archiving *only* electronic versions of their students' theses and dissertations, especially in archiving programs operated by academic libraries. While this steady transition in curatorial practice from print to digital theses and dissertations greatly enhances the current accessibility and sharing of graduate student research, it also raises grave long-term concerns about the potential ephemerality of these digital resources.

Our research focuses on answering the question: *How will institutions address the entire lifecycle of ETDs, ensuring that the electronic theses and dissertations they acquire from students today will be available to future researchers?* We use the phrase *lifecycle management of digital data* in the broad sense defined by the Library of Congress to refer to the "progressive technology and workflow requirements needed to ensure long-term sustainability of and accessibility to digital objects and/or metadata" (Library of Congress, 2006), as well as in the more detailed senses of the digital lifecycle management model as articulated by the Digital Curation Centre in the UK (Higgins, 2008). A key outcome of our research and documentation will be a clearly articulated lifecycle model specific for ETDs.

In order to unpack this complex issue and to assess the library field's ETD lifecycle-management needs and practices, leaders of the Networked Digital Library of Theses and Dissertations (NDLTD) and the MetaArchive Cooperative conducted a series of investigations during 2008-2010. These efforts included surveys, a pilot project, and meetings of the leadership of the two groups, each of which are concerned with different aspects of preserving ETDs. The research team then embarked upon a US Institute for Museum and Library Services-funded project in 2011 to develop guidelines for ETD lifecycle management, software tools to facilitate ETD curation, and educational materials to help prepare ETD curators. As one component of this project, we conducted a focus group with stakeholders. We describe our findings from these surveys below.

1.1 Surveys of ETD Curation Practices

In order to assess practitioner needs and the current status of the field, the MetaArchive Cooperative and the NDLTD conducted a survey in 2007/2008 to examine ETD practices and associated concerns in institutions either currently engaged in ETD programs or considering such preservation service programs. The on-line survey was distributed through five major listservs and received 96 responses, primarily from academic institutions that were providing or strongly considering collection of ETDs and associated ETD services (McMillan, 2008).

Of the survey respondents, 80% accept ETDs, and 40% accept *only* ETDs. The ETD programs report that they accept many formats (more than 20) beyond PDF documents, including images (92%), applications (89%), audio (79%), text (64%) and video (52%). The average size of these programs was 41 GB, and respondents reported 4.5 GB/year average growth. We found that the repository structures used by respondents also vary widely. The more popular approaches included locally developed solutions (34%), DSpace (31%), ETD-db (15%), and such vendor-based repositories as bepress (6%), DigiTool (6%), ProQuest (6%), and CONTENTdm (6%).

This diversity of system types—presumably at least somewhat representative of the overall industry—presents an array of challenges for preservation. Each of these repository systems requires preservation attention during the ingest process to ensure that the materials are submitted in such a way that it is possible to retrieve them and repopulate that repository system with the content. This demands that content carries with it a level of context, and that context differs across repository structures.

The digital collections file and folder structures used by respondents also varied widely. Most respondents reported that their ETD collections are not structured in logically named, manageable virtual clusters. In fact, more than a quarter of respondents reported that their ETD collections are stored in one mass upload directory. This raises many preservation readiness challenges. How can the institution preserve a moving, constantly growing target? How can they ensure that embargoed and non-embargoed materials that often co-exist in the same folder are dealt with appropriately? How will the institution know what these files are if they need to repopulate their repository with them, particularly if they are stored in a repository system that does not elegantly package metadata context with content at export? Only 26% of the institutions manage their ETD collections in annual units. Another 26% use names (departments, authors) or disciplines as unit labels. Seven percent reported using access level labels and another 13% did not know.

The survey also collected information about what information institutions would need to make decisions concerning ETD preservation programs. Perhaps the most remarkable finding from this survey was *that 72% of responding institutions reported that they had no preservation plan for the ETDs they were collecting.*

The responses to this survey led the same researchers to conduct a follow-on survey in 2009 that probed more deeply into digital preservation practices and concerns (Skinner and McMillan, 2009). This survey included questions concerning institutional policies, knowledge and skills needed for digital preservation activities, level of desire for external guidance and expertise in digital preservation, and perceptions about relative threat levels of different factors in the long-term survivability of digital content.

Based on these findings, the MetaArchive Cooperative and the NDLTD undertook a joint pilot project in 2008-2010 to further explore and understand issues highlighted in the surveys and to respond to concerns of their respective memberships about preservation of ETDs. In the course of this pilot project, a group of institutions that are members of both organizations (including Virginia Tech, Rice University, Boston College, and others) worked together to discuss, analyze, and undertake experiments in different aspects of lifecycle management of ETDs, and to identify problem areas experienced by multiple institutions. The pilot project group also explored the literature to better understand what has been published to date on different digital lifecycle management topics, and how such publications relate to ETDs.

During this pilot project, as another means of assessing needs, Gail McMillan (NDLTD) and Martin Halbert (MetaArchive Cooperative) asked a large number of ETD program leaders about their concerns about ETD lifecycle management during workshops conducted at each of three annual ETD conferences hosted by the NDLTD from 2008-2010. Findings from the pilot project analysis and workshop inquiries were reviewed and discussed at three joint planning meetings of the NDLTD board and MetaArchive leadership during this period. They were consistent with the initial findings of the 2007-8 ETD survey.

Similarly, as the *Lifecycle Management for ETDs* project kicked off in 2012, the research team hosted a focus group in conjunction with the February Texas Electronic Theses and Dissertations Association meeting in Denton, Texas. Respondents in this focus group included both College of Arts and Sciences representatives and library representatives. The concerns raised by this group mirrored our earlier findings—most are involved in ETD programs and are either already electronic *only* or will be in the near future. The collection structures, file-types accepted, and repository infrastructures vary wildly. All attendees agreed that establishing documentation, tools, and educational materials that encourage better, more consistent ETD curatorial practices are of great need and should be of value to virtually all categories of academic institutions within the United States and internationally.

2. GUIDANCE DOCUMENTS

There is need for guidance documents in a variety of specific ETD lifecycle management topics to advance the capabilities of institutions that administer ETD service programs. The *Lifecycle Management for ETDs* project has worked to fill these gaps. The research team strongly feels that as a field we need to better understand, document, and address the challenges presented in managing the entire lifecycle of ETDs in order to ensure that colleges and universities have the requisite knowledge to properly curate these new collections. The research team has developed draft documentation on a number of topical areas, as briefly described below.

2.1 Introduction to ETDs

Prepared by Dr. Katherine Skinner and Matt Schultz (Educopia, MetaArchive), this document introduces the “Guidelines” and chronicles the history of ETDs. Using survey data and research findings, it describes the evolving and maturing set of practices in this area. It discusses the philosophical and political issues that arise in this genre of content, including what to do with digitized vs. born-digital objects, how to make decisions about outsourcing, and how to deal with concerns about future publications and

embargoed materials in the lifecycle management framework. The chapter provides a conceptual overview of a lifecycle model for ETDs that makes direct connections between the model and the individual guidance documents described below.

2.2 Access Levels and Embargoes

Prepared by Geneva Henry (Rice University), this document provides information about the ramifications of campus policy decisions for or against different kinds of access restrictions. It defines access restriction and embargo, and discusses reasons for each, including publishing concerns, sensitivity of data, research sponsor restrictions, and patent concerns. It discusses how institutions may provide consistent policies in this area and how policies might impact an institution's lifecycle management practices. It also reviews and compares existing university policies and makes policy recommendations.

2.3 Copyright Issues and Fair Use

Patricia Hswe (Penn State) chronicles ETD copyright and fair use issues that arise both in the retrospective digitization and the born-digital acquisition of theses and dissertations. It discusses institutional stances and guidelines for sponsored research and student work, and also reviews copyright and fair use issues with respect to commercial publishers (including e-book publishers) and vendors such as ProQuest. It seeks to provide clarifying information concerning publisher concerns and issues, providing a concise summary of the relevant information for stakeholders.

2.4 Implementation: Roles & Responsibilities

Xiaocan (Lucy) Wang (Indiana State University) documents the variety of stakeholders who impact and are impacted by the transition to electronic submission, access, and preservation of theses and dissertations, including such internal stakeholders as institutional administration (e.g., president, provost, CIO, general counsel), graduate schools (administrators, students, faculty), libraries (administrators, digital initiatives/systems divisions, technical services, reference), and offices of information technology, and such external stakeholders as commercial vendors/publishers, NDLTD, access harvesters (e.g., OCLC), and digital preservation service providers (e.g., MetaArchive, FCLA, DuraCloud). It emphasizes the range of functions played by these stakeholders in different management phases and institutions.

2.5 Demonstrations of Value

Dr. Yan Han (University of Arizona) provides guidance for institutions concerning assessment of ETD usage, and how communicating such assessment metrics can demonstrate a program's benefits to stakeholders. Han also documents practical examples of documenting and conveying usage metrics for stakeholder audiences, including the university, the students, and the research community more generally. He provides practical guidance for collecting, evaluating, and interpreting usage metrics in support of ETD programs, and discusses how it may be used to refine and promote this collections area.

2.6 Formats and Migration Scenarios

What factors should be considered by colleges and universities to determine what formats they should accept? How can they manage on an ongoing basis the increasingly complex ETDs that are now being produced by students? Bill Donovan (Boston College) discusses these format issues, including "data wrangling" practices for legacy content and migration scenarios for simple and complex digital objects in ETD collections.

2.7 PREMIS Metadata and Lifecycle Events

Another issue revealed in the needs assessment process was that most institutions do not have workflows and systems in place to capture the appropriate levels of metadata needed to manage ETDs over their entire lifecycle. Daniel Alemneh (University of North Texas) informs stakeholders and decision makers about the critical issues to be aware of in gathering and maintaining preservation metadata for ETDs, not just at the point of ingestion, but subsequently, as ETDs often have transitional events in their lifecycle (embargo releases, redactions, etc.). This guidance document will both inform and reinforce the software tools around PREMIS metadata that we are building.

2.8 Cost Estimation and Planning

Gail McMillan (Virginia Tech) provides institutions with information on costs and planning, laying out the critical paths that many ETD programs have charted to date. This document provides cost-benefit analyses of multiple scenarios to give institutions a range of options to consider for their local needs.

2.9 Options for ETD Programs

Our surveys and focus group have demonstrated that many institutions are delayed in ETD program planning simply because they do not have a clear understanding of the range of options to consider in implementing an ETD program. Restricted or open access? Implement an ETD repository or lease a commercial service? Who has responsibility for what functions? Dr. Martin Halbert (University of North Texas) explains the relevant decisions institutions must make as they set up an ETD program and clarifies the pros and cons of different options.

3. LIFECYCLE MANAGEMENT TOOLS

The research team is developing and openly disseminating a set of software tools to address specific needs in managing ETDs throughout their lifecycle. These tools are modular micro-services, i.e. single function standalone services that can be used alone or incorporated into larger repository systems. Micro-services for digital curation functions are a relatively new approach to system integration pioneered by the California Digital Library and the Library of Congress, and subsequently adopted by the University of North Texas, Chronopolis, MetaArchive, Archivematica, and other digital preservation repositories.

The micro-services described below draw upon other existing open source software tools to accomplish their aims. The intent of creating these four micro-services is that they will catalytically enhance existing repository systems being used for ETDs, which often lack simple mechanisms for these functions.

3.1 ETD Format Recognition Service

Accurate identification of ETD component format types is an important step in the ingestion process, especially as ETDs become more complex. This micro-service will: 1) Enable batch identification of ETD files through integration of function calls from the JHOVE2 and DROID format identification toolkits; and 2) Structure micro-service output in ad hoc tabular formats for importation into repository systems used for ETDs such as DSpace, and the ETD-db software, as well preservation repository software such as iRODS and DAITSS and preservation network software such as LOCKSS.

Components & Basic Requirements:

JHOVE2, DROID, XML output schema, Utility scripts (run commands, output parsers, etc.) & code libraries, API function calls, System requirements, Documentation & instructions

3.2 PREMIS Metadata Event Record-keeping

One gap highlighted in the needs analysis was the lack of simple PREMIS metadata and event record keeping tools for ETDs. This micro-service needs to: 1) Generate PREMIS Event semantic units to track a set of transitions in the lifecycle of particular ETDs using parameter calls to the micro-service; and 2) Provide profile conformance options and documentation on how to use the metadata in different ETD repository systems.

Components & Basic Requirements:

PREMIS Event profiles (example records) for ETDs, Event-type identifier schemes and authority control, AtomPub service document & feed elements, Utility scripts (modules) & code libraries, API function calls, Simple database schema & config, System requirements, Documentation

3.3 Virus Checking

Virus checking is an obvious service needed in ETD programs, as students' work is often infected unintentionally with computer viruses. This micro-service will: 1) Provide the capability to check ETD component files using the ClamAV open source email gateway virus checking software; 2) Record results of scans using the PREMIS metadata event tracking service; and 3) Be designed such that other anti-virus tools can be called with it.

Components & Basic Requirements:

ClamAV, Utility scripts (run commands, output parser, etc.) & code libraries, API function calls, System requirements, Documentation & instructions

3.4 Digital Drop Box with Metadata Submission Functionality

This micro-service addresses a frequently sought function to provide a simple capability for users to deposit ETDs into a remote location via a webform that gathers requisite submission information requested by the ETD program. The submission information will: 1) Generate PREMIS metadata for the ETD files deposited; 2) Have the capacity to replicate the deposited content securely upon ingest into additional locations by calling other Unix tools such as rsync; and 3) Record this replication in the PREMIS metadata.

Components & Basic Requirements:

Metadata submission profile(s), Client/server architecture, GUI interface, SSL, authentication support, Versioning support, Various executables, scripts & code libraries, Database schema & config, System requirements, Documentation

All of these tools will be documented and released in 2013 via the project site: <http://metaarchive.org/imls>.

4. CONCLUSIONS

The first phase of this project has helped to reinforce preliminary research we had conducted regarding ETD lifecycle management practices (or the significant lack thereof). The field has a dire need

for descriptive, not prescriptive, documentation regarding the range of ETD programs that institutions have designed and implemented to date, and the variety of philosophical, organizational, technical, and legal issues that are embedded therein. The field also has a stated need for lightweight tools that can be quickly implemented in a range of production environments to assist with some of the commonly needed curatorial practices for lifecycle management of these collections.

5. ACKNOWLEDGMENTS

We greatly appreciate the generous support of the Institute for Museum and Library Services (IMLS).

6. REFERENCES

- Caplan, Priscilla. "The Preservation of Digital Materials." *Library Technology Reports*, (2008) 44, no. 2.
- Conway, Paul. "Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas." *Library Quarterly*, (2010) 80:1, 61-79.
- Fox, Edward A., Shahrooz Feizabadi, Joseph M. Moxley, and Christian R. Weisser, eds. *Electronic Theses and Dissertations: A Sourcebook for Educators, Students, and Librarians*. New York: Marcel Dekker, 2004.
- Halbert, Martin, Katherine Skinner and Gail McMillan. "Avoiding the Calf-Path: Digital Preservation Readiness for Growing Collections and Distributed Preservation Networks," *Archiving 2009*, Arlington, VA, May 2009, p. 86-91.
- Halbert, Martin, Katherine Skinner and Gail McMillan. "Getting ETDs off the Calf-Path" ETD 2009: *Bridging the Knowledge Divide*, Pittsburgh, PA, June 10-13, 2009. Sharon Reeves, ed. <http://conferences.library.pitt.edu/ocs/viewabstract.php?id=733&cf=7>
- Hall, Susan L., Lona Hoover, and Robert E. Wolverton, Jr.. "Administration of Electronic Theses/Dissertations Programs: A Survey of U.S. Institutions." *Technical Services Quarterly* 22, no. 3 (2005): 1-17.
- Lippincott, Joan K. "Institutional Strategies and Policies for Electronic Theses and Dissertations." *EDUCAUSE Center for Applied Research Bulletin*, no. 13 (2006). <http://net.educause.edu/ir/library/pdf/ERB0613.pdf>
- Lippincott, Joan K., and Clifford A. Lynch. "ETDs and Graduate Education: Programs and Prospects." *Research Library Issues*, no. 270 (June 2010): 6-15. <http://publications.arl.org/rli270/>
- McMillan, Gail. "ETD Preservation Survey Results." *Proceedings of the 11th International Symposium on ETDs*, Robert Gordon University, Aberdeen, Scotland. (June 2008) <http://scholar.lib.vt.edu/staff/gailmac/ETDs2008PreservPaper.pdf>
- McMillan, Gail, and Katherine Skinner. (2010) "NDLTD/MetaArchive Preservation Strategy." (3rd ed.) <http://scholar.lib.vt.edu/theses/preservation/NDLTDPreservationPlan2010.pdf>
- Skinner, Katherine, and Gail McMillan. "Surveys of Digital Preservation Practices and Priorities in Cultural Memory Organizations." 2009 NDIIPP Partners Meeting, Washington, D.C., June 24, 2009. http://www.digitalpreservation.gov/news/events/ndiipp_meetings_ndiipp09/docs/June24/NDIIPP_Partners_2009_finalRev2.ppt

Preservation Watch: What to monitor and how

Christoph Becker,
Kresimir Duretec,
Petar Petrov
Vienna University of
Technology
Vienna, Austria
{becker,duretec,petrov}
@ifs.tuwien.ac.at

Luis Faria,
Miguel Ferreira
KEEP SOLUTIONS
Braga, Portugal
{lfaria,mferreira}@keep.pt

Jose Carlos Ramalho
University of Minho
Braga, Portugal
jcr@di.uminho.pt

ABSTRACT

For successful preservation operations, a preservation system needs to be capable of monitoring compliance of preservation operations to specifications, alignment of these operations with the organisation's preservation objectives, and associated risks and opportunities. This requires linking a number of diverse information sources and specifying complex conditions. For example, the content to be preserved needs to be related to specifications of significant properties, to the file formats used to represent that content, to the software environments available to analyse, render and convert it and the technical environments available to the designated user communities that will consume this content.

This article analyses aspects of interest in a preservation context that call for automated monitoring and investigates the feasibility of drawing sufficient information from diverse sources together and linking it in a meaningful way. We define a number of preservation triggers that lead to preservation planning activities and present the requirements and a high-level design of a preservation watch system that is currently being developed. We demonstrate the value of such a monitoring approach on a number of scenarios and cases.

Categories and Subject Descriptors

H.1 [Information Systems]: Models and Principles; H.3 [Information Systems]: Information Storage and Retrieval; H.3.7 [Information Systems]: Information Storage and RetrievalDigital Libraries; K.6.4 [Computing Milieux]: Management of computing and Information Systems—*System Management*

Keywords

Digital Preservation, Preservation Planning, Monitoring, Watch

1. INTRODUCTION

Digital Preservation is in essence driven by change of organisational and technical kind. Aspects of change range from platform technologies and rendering environments to storage media, shifting modes of access and interactivity, and finally, shifts in the semantics of information itself. Any archival information system thus needs to continuously adapt to changing environments to ensure alignment between preservation operations and the goals and objectives of the system.

Monitoring is recognised as a key element of successful preservation. However, to date it is mostly a manual process that is sporadically initiated as a reaction to urgent questions. At best, technical reports are produced about selected topics and circulated within the community.

More and more cases of successful preservation operations are being developed. The SCAPE project¹ is focusing on scalable operations for preserving massive amounts of information through data-centric parallel execution mechanisms [7]. However, for such operations to be scalable and successful over time, automated mechanisms and processes for control and monitoring need to be designed.

Isolated strands of systematically collecting information that can be used to guide preservation decision making have been developed. Well-known examples include registries of file formats or emulation environments. However, these are far from being complete in the information they cover, and there are few links between the islands of information.

For an organisation responsible for managing a digital repository over time, the corresponding *monitoring capabilities* that are required can be described as

1. *Internal monitoring* of the systems in place, the operations in place, the assets and activities, and
2. *External monitoring* of the world of interest such as user communities, technologies, and available solutions.

Based on these systematic information gathering processes, preservation planning as decision making capability can then act well-informed to ensure that what the organisation does to keep content authentic and understandable is sufficient and optimal. A number of questions arise in this context.

1. Which are the key aspects that need to be monitored? What are the main entities and which properties need to be considered?
2. How can the information be collected? How can it be represented?

¹<http://www.scape-project.eu>

3. How can the information be linked together so that connections between parts of the data can be made?
4. What properties does a system need to possess to enable automated monitoring and linkage between the relevant aspects?
5. How can this be deployed and used in a way that multiple organisations can mutually benefit from each other's effort and experience?
6. How can we ensure that such a platform will be extensible and create synergies between different players so that the knowledge base grows continuously?

In this article, we discuss these questions and propose a design for such a system. We envision a ‘Watch component’ to collect information from a number of sources, link it appropriately together, and provide notifications to interested parties when specified conditions are satisfied. The motivation of this Watch component is in part driven by the experience gathered in preservation planning: The preservation planning tool Plato² provides powerful and systematic decision making support and includes an increasingly formalised model of relevant aspects, entities and properties to be considered in preservation planning [9]. Plato is not designed to provide continuous monitoring capabilities, but preservation plans specified as the result of such planning activities specify which aspects to monitor based on the influencers considered in decision making. It is then the responsibility of the Watch component presented here to continuously monitor the state of the world and of the system in question to determine whether conditions are met that may require an update of plans and operations.

Section 2 outlines the background of monitoring in the context of preservation and illustrates typical information sources that are of relevance. In Section 3 we discuss a model of drivers, events, conditions and triggers for preservation watch. Section 4 describes the various sources of information that are being leveraged. Section 5 summarises the key design goals of a preservation watch system and presents a high-level design of the Watch component. Section 6 illustrates typical conditions and benefits of the approach in a scenario based on a real-world preservation planning case, while Section 7 summarises key risks and benefits and outlines the next steps ahead.

2. BACKGROUND

Monitoring is a common subject in any domain that must cope with the demands of a changing environment. It is a major input for decision-making and ensures that specified plans are continuously adapted to changes in the environment. Monitoring feeds back to decision-making to close a continuous adaptative cycle [4]. In the digital preservation domain, monitoring is especially critical as the domain challenge itself stems largely from rapidly changing environments. The need and concept of monitoring have been identified and discussed before [5, 2]. However, these all focus on a very high-level approach of preservation monitoring and do not define a systematic method or guideline on how to accomplish that capability. Furthermore, the tools presently known to support preservation monitoring are mainly manual, incomplete and used in an ad-hoc fashion. Monitoring now comes in the form of research studies and technical

²<http://www.ifs.tuwien.ac.at/dp/plato>

reports, format and tool registers, and generic application catalogues outside of the core preservation domain [8].

An early influential report on file format risks and migration strategies discusses risks that executing or postponing a migration might introduce [13]. Regular Technology Watch Reports of the Digital Preservation Coalition provide focused discussions on emerging topics and include technical investigations.³ However, none of this is machine-understandable.

Online registries with technical information about file formats, software products and other technical components relevant to preservation have been available for some time. This includes the well-known examples PRONOM⁴, The Global Digital Format Registry⁵ (GDFR) [1], and the newly released Unified Digital Format Registry⁶ (UDFR). Complementary approaches include the P2 registry⁷ based on semantic web technologies [15], and the Conversion Software Registry⁸. Unfortunately, these online registries are not yet functioning or are not very complete. For example, relevant risk factors *per format* are only covered for a handful of entries.

Online software catalogues monitor new versions of software for a generic domain use. These sites do not specifically consider digital preservation aspects, but provide comprehensive descriptions and commonly have a social component that can contain interesting information. Some examples of these sites are CNET’s download.com⁹ and [iUseThis](http://iusethis.com)¹⁰. App stores like Apple’s Mac App Store and Ubuntu’s Software Center and repositories can also be a good source of information. In the domain of digital preservation, TOTEM - the Trustworthy Online Technical Environment Metadata Database tries to address the gap of linking environments and compatible software, but is limited to emulated environments addressed within the KEEP project.¹¹

This overview relates solely to file formats and tools for conversion of file formats or emulation, but many more information sources are required. Furthermore, it has to be noted that sources that focus on digital preservation have a generally very reduced coverage (registries) or machine-readability (reports), while general purpose sources normally cover very limited facets of the information relevant for digital preservation. Finally, none of these sources allows preservation monitoring to be done automatically and alert the user when a preservation risk is identified. However, this step towards automation is crucial: As content grows in volume and becomes increasingly heterogeneous, the aspects of technologies that need to be monitored are by far outgrowing any organisation’s manual capabilities.

The OAIS model [5] includes, within the functional entity Preservation Planning, the functional components “Monitor Designated Community” and “Monitor Technology”. These provide the core monitoring functions in a repository scenario. Monitoring the user community is meant to focus

³<http://www.dpconline.org/advice/technology-watch-reports>

⁴<http://www.nationalarchives.gov.uk/PRONOM/>

⁵<http://www.gdfr.info>

⁶<http://www.udfr.org>

⁷<http://p2-registry.ecs.soton.ac.uk>

⁸<http://isda.ncsa.uiuc.edu/NARA/CSR>

⁹<http://download.com>

¹⁰<http://iusethis.com>

¹¹<http://keep-totem.co.uk>

on “service requirements and available product technologies”, while technology monitoring includes “tracking emerging digital technologies, information standards and computing platforms (i.e., hardware and software)” [5]. The OAIS also mentions monitoring functions of the Administration entity as well as monitoring archival storage and systems configurations, but these are seen as separate functions.

Historically, the identification of risks by monitoring tools has been delegated into other tools such as file format registries or seen as a manual task such as providing technical reports. A step forward in automation was done in the initiative to create an Automatic Obsolescence Notification Service (AONS) that would provide a service for users to automatically monitor the status of file formats in their repositories against generic format risks gathered in external registries and receive notifications [14]. The system would gather collection profiles from repositories, by using format identification tools on content, and seek obsolescence risk indicators on file format information in external registries (like PRONOM). The system would also allow caching and extending format registries and the creation of new adaptors for new registries. The notification service allows subscription of various events, like end of a repository crawl or change in the information about a format, and send a notification via email, RSS feed and task boxes on the GUI.

AONS was limited to gathering information about file formats, assuming that all other aspects of the world that would be relevant for digital preservation would be gathered, assessed, and represented in a structured way by the external registries. This assumption and the lack of available information in format registries constrained the usefulness of the system. Moreover, not all desired features defined in the concepts could be successfully completed.

The lack of a defined methodology for systematic preservation monitoring and tools that help to enforce and automatize this capability forces content holder institutions to create their own methodology, highly based on manual effort and therefore scattered throughout many departments and different people via responsibility distribution, which results in a partial and stratified view of the properties that condition decisions. Furthermore, most institutions cannot afford the effort to have even this partial view on the environment and thus ignore or postpone efforts for preservation monitoring [6].

In contrast, a well-designed monitoring system would inspect the properties of the world and provide needed knowledge to identify risks. This knowledge requires an integration of several aspects of the world – tools and formats, but also content, empirical evidence, organizational context, technology, user trends, and other aspects. Furthermore, this information needs to be cross-referenced and analytically accessible for automated procedures so that indications of risks and opportunities can be found and deep analysis processes (such as a manual intervention) can be initiated only when needed.

Our investigation of the question *What to monitor?* can build on a number of additional analytical steps that have been taken previously. On the one hand, the Plato preservation planning framework provides a systematic guidance on analysing influence factors. On the other hand, systems-oriented approaches such as SHAMAN provide a categorisation of drivers and constraints [2]. Table 2 classifies key drivers in a DP scenario in internal and external categories.

Table 1: DP drivers according to SHAMAN [2]

Internal	
Business Vision	Goals, Scope of designated community, etc.
Resources	Infrastructure (e.g., operational costs, expertise needed), Hardware (e.g., operational costs, technological capability), Software (e.g., operational costs, technological capability), Staff (e.g., expertise and qualifications, commitment)
Data	Volume, Structure, Representation, Semantics, etc.
Processes	Dependencies, Responsibilities, Alignment, etc.
External	
Producers	Demand satisfactions, Content, Technology, Trust and reputation
User community	Technology, Knowledge, Demand satisfaction, Trust and reputation
Contracts	Deposit, Supplier and service, Interoperability, Access, etc.
Supply	Technology, Services, People
Competition	Overlap of: Services, Content, User community, Producers, Technology, Mandate, Rights, Funding, Capabilities
Regulation and mandate	Regulation/Legal constraints, Embedding organization regulation, Mandate, Rights and ownership, Certification, Funding

Any of these drivers feature conditions that influence decisions and operations for preservation. Section 4 will discuss which information sources we can connect to for gathering information about these drivers.

3. AUTOMATED PRESERVATION WATCH

We envision a *Preservation Watch component* as a system that enables automated monitoring of operational preservation compliance, risks and opportunities by collecting, fusing and analysing information from various sources. From an abstract perspective, the usage of such a Watch component can be reduced to the following steps:

1. An actor has a question about a certain aspect of the world that is of interest to the agent.
2. The actor expresses this interest in the form of a question about a property that represents this interest.
3. The function of Watch then is to find a method to deliver an answer to this question that is timely and reliable.
4. Having received an answer to the question, the actor will want to assess the meaning and impact of this answer. This may require consultation of a decision-making capability.

The relevant aspects of the world, i.e. the entities and their properties about which information should be gathered, are expected to evolve and expand over time. The initial model is focused on the core question of information representation, formats and available environments. These are not the only sources of information, but instead represent the seed of key drivers to be considered.

Figure 1 shows a minimal model of the main entities of interest in the initial phase of Watch. Each of the entities shown has a number of known and named properties of interest. A few additional relationships and entities are omitted here for clarity. Organisations holding ownership

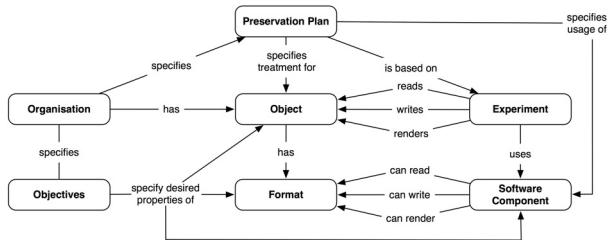


Figure 1: Minimal initial domain– model

and responsibility of objects specify objectives that relate to these objects, the resources required to preserve them, the processes used to preserve and access them and the software components used to run these processes. Objects have a number of properties, including the significant properties that need to be kept unchanged to preserve the authenticity of the object. The format of the representation is a key property and itself has a number of properties, such as the well-known risk factors commonly used to assess formats for their risk and benefits in preservation [15].

Software components for analysing, migrating, rendering and quality assuring content are key elements to ensure continued access. They are evaluated systematically in controlled experiments to provide the evidence base for the decisions specified in preservation plans [3]. These plans specify particular preservation actions to treat content for keeping it understandable and authentic. Systematic experiments are executed to test certain properties of software components on certain test data sets, containing objects. With increasing automation, systematic experiments can be scaled up and systematically conducted on large volumes of content [7]. Considering that such experiments are increasingly common across the DP domain, analysing these can uncover hidden risks and opportunities for operations in related scenarios.

Linking such information *across scenarios* enables us to answer critical questions such as the following.

- How many organisations have content in format X?
- Which software components have been tested successfully in analysing objects in format Y?
- Is the Quality Assurance tool Q, which checks documents for the equivalence of their textual content, reliable for this pair of formats?
- Has anyone encountered problems in terms of stability for this migration component when executed on very large image files?

The questions outlined above refer to known properties of identified classes such as *software components* and their properties [9]. As the preservation community becomes aware of this platform and the mutual benefits to be had from synergistic data collection, we expect this minimal model above to evolve substantially and cover additional entities and properties of interest.

For the recipient, an answer may have a variety of impacts and can overlap with other answers. Decisions are generally taken with a number of influencing factors in mind: For example, a decision to postpone a migration project may be driven by considerations on migration costs, the availability of automated tools to perform quality assurance on conversion processes, storage costs and cost models, and specific

rendering environments that are at this point in time able to support content delivery to the user communities [12]. Over time, these drivers can change simultaneously. Each change can be critical, but it is only considering all relevant aspects that informed decisions can be taken.

This means that there may be simple conditions attached to a question. These conditions trigger an event when they are met, for example when the answer changes by more than 5%. The role of an automated watch process is not to assess the cumulative impact of multiple answers and what meaning they have to an external consumer of the answers. Essentially, the Watch component itself should be agnostic of the ultimate effects of changes: Its primary purpose is to make the state of the world available for assessment, not to assess it.

4. SOURCES OF INFORMATION

For any given question, several sources of information will often have to be consulted. This section gives an overview of possible sources in terms of the information they provide and attempts a high-level categorization.

Content profiles. A content profile provides statistical data about digital content of any type and offers an aggregated view of content based on its metadata, in particular detailed technical characteristics. An organisation's own content profile thus provides the basis for in-depth analysis and risk assessment. The quality of any such analysis depends on the richness of information present. While the formats contained in a repository are the first property that comes to mind, it is critical to perform a deeper analysis on other properties to uncover dependencies, feature distributions and hidden risks. By linking information such as the presence of content-specific features, embedded content types or other aspects such as the presence of digital rights management, it becomes possible to monitor often-overlooked preservation issues and collect information that can be meaningfully shared even across organisations.

An entirely different aspect can be covered when considering *others'* content profiles and content profiling on large-scale public content such as web archives. Given the massive data volumes presented there, in-depth profiling of content over time would allow us to provide indicators for file format adoption and impending obsolescence. Specific content profiles of comparable organisations, on the other hand, can enable risk assessment and comparison as well as facilitate collaboration.

Format registries. Changes in the properties of existing formats or the appearance of new formats need to be detected and compared with organisational risk profiles and content profiles. Examples of this type of sources are the PRONOM and P2 registries. However, the crucial point is the coverage of information which current information sources are still severely lacking. Designs for these systems have traditionally relied on inherently closed-world models. Moderated registries such as PRONOM have not shown to be very responsive in capturing the evolving knowledge that is available. The P2 format registry showed the benefits of Linked Data for such format information [15], and increasingly, open information models using RDF and ontologies are leveraged to capture the inherently evolving nature of format properties. This semantic web approach makes efforts such as the new UDFR building on OntoWiki a potentially very valuable source.

Software catalogues. Software components for identification, migration, characterisation or emulation are at the heart of preservation operations. We broadly categorise preservation components into Action, Characterisation and Quality Assurance components. *Action* components perform operations on content or environments, such as migration and emulation. *Analysis* components provide measures of properties in content, such as a format identification or the presence of encryption or compression. *Quality Assurance* components, finally, perform QA on preservation actions, such as algorithmic comparisons of original and converted objects or run-time analysis of rendering quality.

Table 4 lists exemplary change events that can be triggers for preservation activities. Components are continuously developed: New components are published and new versions of components are developed. These components might provide new and better migration paths, new options for performing Quality Assurance, or new and better opportunities for analysing existing content. On the other hand, new knowledge about existing components is gained continuously and, when shared, can provide tremendous value to the community.

Experiments. The role of evidence is central to trustworthy Digital Preservation [?]. In addition to collecting declared published information from catalogues, empirical evidence from controlled experiments are a valuable source of information. On the one hand, preservation planning experiments are executed on a subset of a collection to provide manually validated, deep insights into potential alternative actions [3]. These experiments provide valuable knowledge not only for the planning scenario in question but also for future usage. They are executed only on a subset of a whole collection, but processing this subset can still take a significant amount of time. Moreover, the experiment results will often be validated and amended manually and are therefore particularly valuable. Publishing such experimental data so that the results can be accessed can provide significant benefits [11]. On a much larger scale, the Linked Data Simple Storage specification (LDS3)¹² is being positioned to enable large-scale publication of Linked Data sets in digital preservation, describing content statistics, experimental results in content validation and conversion, benchmarks, and other experimental data. This can be used to publish experimental data from any platform and environment, as long as it is properly described.

We note that the combination of the above three information sources goes a long way in answering the questions outlined in Section 3. However, they do not cover the questions of internal systems monitoring and the alignment between a preservation system and its objectives. These are covered by the following sources.

Repository systems. Two aspects about repositories are considered: On the one hand, the state of a repository and the content it is holding is of interest (What is the growth rate of content? Are all objects covered by preservation plans?). On the other hand, repositories perform continuous operations that can provide valuable information to feed into decision making. This includes validity check as well ingest and access operations (What is the average access time using migration upon access? How many access requests have failed?) By specifying a standardised vocab-

Table 2: Examples of software triggers

Event	Example Cause
New software	New migration software for specific formats used in the repository
	New analysis or characterization software for a certain content type or format
	New QA software for a certain content type
	New monitoring service for a question of interest
	New software version release
New knowledge about software	New testing results for a action, analysis, quality assurance or monitoring software used in the content repository
	Change in software dependencies
New repository system or version	New software release, acquisition of a new system
Capabilities	Optimized internal infrastructure leads to new technical opportunities (e.g. faster throughput in organizational SOA)

ulary for the core set of such events, it becomes possible to monitor whether the performed preservation activities are successful. This also supports anticipation of trends in the repository operations and usage.

Organisational objectives. Changes in the organisations strategies and goals may respond to shifts in regulations or changes in priorities. These high-level elements of governance will be reflected in the policies of an organisation and ultimately in the specific objectives for preservation. If such objectives can be formalised, they will provide a critical starting point for monitoring fulfilment of these objectives on specified indicators. Ongoing work is formalising such a model using semantic web technologies. This can be fed into a knowledge base to enable direct queries resolving objectives against the state of the world.

Simulation. Based on trends that can be inferred from gathered data, models can be created to predict future events and the consequences of preservation actions on repositories and other preservation environments [16]. These predictions can be fed back into a watch system as a source of information, allowing the detection of possible risks before they actually happen. This is especially important for large-scale repositories where understanding of storage and computational resources is crucial.

Human knowledge. Finally, human users should be able to insert information about every possible entity (objects, format, tool, experiment, repository status, etc.).

All these sources will evolve and change. They can cease to exist, modify their behaviour or the way information is published, even the type of information or the way it is structured. The monitoring system should be designed to allow for this through a loosely coupled architecture of information adaptors. It allows the update, addition and replacement of sources, so that the configuration of complementary information sources can evolve over time.

We observe that these sources differ in their structure, ranging from highly structured linked databases to operational systems that raise events through log mechanisms. Furthermore, some of these drivers are internal, such as the operations specified by plans and the operational attributes of the system, while others are external. Attributes of the surrounding environment can influence plans, policies and operations [2]. Some sources can be both internal and external, since information internal for one organisation can

¹²<http://www.lds3.org>

(in anonymised form) be of tremendous interest to another. For example, the format profile of a repository is internal to the organisation, but when shared, it can be used to assess whether a format is commonly used and can be considered a *de facto* standard.

5. A MONITORING SYSTEM

To collect, link and analyse information in the way described, a Watch component should aim for the following high-level goals.

1. **Enable a planning component such as Plato to automatically monitor entities and properties of interest.** Plans are created based on an evaluation of specific alternatives against formally modelled criteria [9]. A plan can thus cause a number of questions and conditions that can be tracked continuously to verify the compliance of operations to plans and detect associated risks and opportunities. The Watch component shall enable this tracking and support round-trip evolution of plans. We will discuss this in Section 6.
2. **Enable human users and software components to pose questions about entities and properties of interest.** Components and human users will be able to pose questions to the Watch component and receive answers about the measures. They can also deposit conditions to receive a notification upon significant changes.
3. **Collect information from different sources through adaptors.** Different sources will be relevant for the Watch component. Each source of information provides specific knowledge in different information models that will have to be mapped, normalized, merged and linked.
4. **Act as a central place for collecting relevant knowledge that could be used to preserve an object or a collection.** Information for object/collection preservation shall be collected and linked so that the Watch component provides a uniform reference point for gathering information about a variety of aspects.
5. **Enable human users to add specific knowledge.** While growing automation enables scalable data collection, human input is and will remain a valuable source of information.
6. **Notify interested agents when an important event occurs** through configurable notification channels.
7. **Act as an extensible platform.** This last item is particularly important: The Watch component is intended to function as a platform on which additional information sources can be added and connected easily.

Figure 2 shows the main building blocks of the Watch component, which is currently under development ¹³. A number of external sources are monitored through a family of *adaptors* as outlined in Section 4. These correspond to an adaptor interface and deliver measures of defined and named properties of interest. A set of interfaces is defined to allow pulling information from the outside world, specifically used when the relevant sources remain agnostic to the

¹³<https://github.com/openplanets/scape-pw>

watch service. These also serve to push information into the Watch component, used for example when sources of information need more control over what information is sent and how. Finally, a manual web user interface empowers users to send information when no automatic source is available. The extension of additional adaptors is supported by a dynamic plug-in architecture that relies on automated discovery of applicable information sources based on the questions posed.

Several adaptors can be linked together through a Monitoring Service which configures each adaptor and delegates information collection. Such adaptors extract information from a particular source, analyse and transform the information model if necessary and provide measures of specified properties of interest in a well-defined information model. The monitoring services can thus feed the collected information into the knowledge base.

The knowledge base presents a generic data model that is able to capture different measurements of relevant properties for digital preservation in an internal Linked Data store [10]. The data within the store represent a focused, curated part of preservation-related properties of the world. This can be used by queries and analyzed for occurrences of relevant properties, events and anomalies. All internal components make use of this internal model, which specifies a common language for data exchange [6].

To enable the structuring of information, the knowledge base data model must define two sets of elements. One, managed administratively, describes which model of the world is covered and accepted by defining which types of entities can exist and which properties are defined for them. Another, managed by the sources of information, describes instances of these entities and the values of their properties. The data model also keeps a register of the information provenance and history of changes to allow the traceability of information, which will transitively improve the traceability of the decision making process.

A key design decision concerns the representation of the collected information and how it should be linked. It is clear that the data layer of the Watch component must support the flexible model described for the knowledge base and provide the reasoning and querying features needed to answer complex questions about the world. The most appropriate technology for the representation of this model is clearly Linked Data implemented as a Triplestore. Relying on semantic models has strong benefits: It allows flexible integration and extension of the data model, while at the same time supporting inference and reasoning and ensures the extensibility of the underlying ontologies. An in-depth analysis of scalability issues of Linked Data concluded that the current state of the art in query performance is more than adequate for the estimated volume and throughput needed by the Watch component data layer [6].

6. MONITORING SCENARIOS AND EVENTS

To illustrate the effect of creating a component that addresses the goals envisioned, consider an organisation running a large digital repository. The organisation has connected the repository to a central deployment of the component (or created its own deployment based on the openly available code base). We can assume that the Watch component constantly monitors sources like format registries and component catalogues allowing users to pose questions about

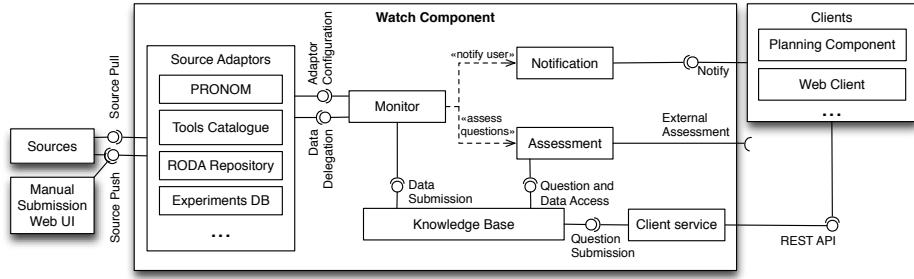


Figure 2: High-level architecture of the Watch component

different format or component properties.

There are two primary situations when a user would pose a question to the Watch component.

First, a typical case is the completion of a preservation plan as described in [3] and illustrated in [12]. A plan specifies a number of decision criteria, including format objectives, preservation action constraints and goals, and metrics specifying how to verify and validate authenticity of content upon execution of a preservation plan (such as a large-scale migration). The plan selects and specifies the best-performing action based on its real performance against these criteria. Upon deployment and operational execution of that plan, the organisation needs to verify that large-scale operations perform according to specifications (operational compliance). It also benefits from automated monitoring of potential risks and opportunities. For example, a large-scale experiment conducted by another organisation on content in the same format may uncover a systematic bias of a quality assurance tool when measuring image quality in TIFF-to-JP2 conversions with embedded color profiles. Such a bias is a potential risk to the authenticity of content that should be raised by a monitoring system. On the other hand, risk factors of format will change over time and should lead to appropriate events.

Second, a key feature of the proposed design is the fact that organisational policies and objectives can be linked to the information collected in monitoring, so that automated queries are able to resolve the question to which degree the current state of a repository matches the goals of an organisation. Hence, a set of standard conditions can be activated to perform such automated monitoring. As soon as organisational objectives change and are fed into the knowledge base through an appropriate adaptor, an automated compliance check can be performed.

What happens now in the event of a change detected by this monitoring process? As outlined earlier, an event raised by the Watch component can be assessed for its significance internally, according to its conditions, and externally, in its contextualised impact on preservation operations. Such an external assessment may consist in recalculating scores for alternative preservation strategies with updated information, which is fully supported by the utility approach followed by Plato [3]. If the assessment concludes that a change in operations is advisable, an iteration of the planning workflow will lead to a revision of the corresponding preservation plan and an update of the associated monitoring conditions. This leads to a continuous monitoring lifecycle of evolving

plans and operations.

Table 3 summarizes these and other potential events. While space prohibits an in-depth discussion of all drivers, indicators and conditions, it can be seen that the fusion and interlinking of such diverse, yet related sources provides a powerful mechanism for monitoring. An in-depth discussion on these triggers and a full data model for specifying questions, conditions and triggers can be found in [6].

7. DISCUSSION AND OUTLOOK

Monitoring the preservation environment is a crucial part of the long-term viability of a system and the data it preserves. Monitoring supports the planning process with continuous tracking of the suitability of the decisions, delegating the risk assessment of the perceived significant changes back to planning. Such a monitoring system is essential for continued digital preservation.

So far, manual processes are used to track all the environment variables that might affect the multitude of object file formats within a repository, with all their different characteristics and contexts. Currently, no tool or service exists that could properly provide this function in a scalable way.

This document delineates the design and development of such a system, named the Watch component, based on the knowledge of past research and experience and going forward by defining new concepts and strategies. Based on real-world scenarios and an analysis of drivers and possible sources of information, we outlined the key sources to be monitored and specified the requirements and high-level design of a Watch component that is currently under development. This architecture enables organisational capability development through flexible and extensible services. The presented design supports automated elements of preservation management without constraining organisations to follow a specific model in their deployment of the planning capabilities. By modelling and implementing watch mechanisms, triggers, and suitable actions to be taken for each trigger, this system supports closed-loop preservation processes in which automated monitoring of collections, actions, plans, systems, and the environment triggers appropriate diagnosis and reaction.

Current work is focused on developing and connecting information source adaptors and providing API specifications that allow additional adaptors to be connected easily. Furthermore, the planning component Plato is being extended to support evolving lifecycles of preservation plans and provide automated assessment of accumulated changes against organisational objectives.

Table 3: From drivers to information sources: Exemplary preservation triggers[6]

Driver	Questions	Indicators	Example conditions	Sources
Content	Is there corrupted content?	Completeness validation fails Access fails	Log contains validation failure Access failure event reported	Repository Repository, User
	Is any content being ingested into the repository?	Ingest activity notices new content	Format of ingested content is different from content profile	Repository, Ingest, Collection profiler
	Is the content volume growing unexpectedly?	Rate of growth changes drastically in ingest	Growth of collection X exceeds threshold	Repository, Ingest, Collection profiler
	Which text formats appear in collection X?	New acquisition activity, new ingest activity	Mode (format) changes	Collection profile, Repository
Operations	Do we have plans defined for all collections?	Mismatch between content profile and set of plans	Exists collection with size greater than threshold defined in policy model without a plan	Repository, plans, policies
	Are our content profiles policy-compliant?	Mismatch between content profile and format/representation objectives	Content exhibits properties that are not acceptable (e.g. encryption)	Content profiles, policy model
Producers and Consumers	Are there any new or different producers?	New producer uses ingest process, new producers send new material	Exists new producer	Repository
	Are there any new or changed consumers?	New consumers use access process	Exists new consumer	Repository
Policies	What is the current valuation of collection X?	Change in policy model	Valuation changes	Policy model
Software	Which experiments have tested migration from TIFF to JPEG2000?	Evaluation data in an experiment platform	Exists new experiment with specified entities and properties	Experiment results
	Is there any software for converting TIFF to JPEG2000 that is applicable to our server environment?	New software matching certain criteria is tested within the experiment database	Matching migration component is tested on >10K objects on Linux platform without crashes or corrupt output	Experiment results, Component catalogue
Storage	What will be the content volume in 18 months? What will be the monthly storage costs?	Simulator prediction	In time X, the total size will be above a certain threshold	Simulator
Format	What is the predicted lifespan of format X?	Simulator prediction	The obsolescence time is below threshold	Simulator
New format risk	What is the risk status for the format X	New risk defined in the policy model	There is a new control policy about required format properties	Policy model

Acknowledgements

Part of this work was supported by the European Union in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

8. REFERENCES

- [1] S. L. Abrams. Establishing a global digital format registry. *Library Trends* 54 (1) Summer 2005, pages 125–143, 2005.
- [2] G. Antunes, J. Barateiro, C. Becker, J. Borbinha, D. Proen  a, and R. Vieira. Shaman reference architecture (version 3.0). Technical report, SHAMAN Project, 2011.
- [3] C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries*, 10(4):133–157, 2009.
- [4] S. Beer. *Brain of the Firm*, volume 1st ed. John Wiley & Sons, 1981.
- [5] Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1, 2002.
- [6] K. Duretec, L. Faria, P. Petrov, and C. Becker. *Identification of triggers and preservation Watch component architecture, subcomponents and data model*. SCAPE D12.1, 2012.
- [7] O. Edelstein, M. Factor, R. King, T. Risse, E. Salant, and P. Taylor. Evolving domains, problems and solutions for long term digital preservation. In *Proc. of iPRES 2011*, 2011.
- [8] M. Ferreira, A. A. Baptista, and J. C. Ramalho. A foundation for automatic digital preservation. (48), July 2006.
- [9] M. Hamm and C. Becker. Impact assessment of decision criteria in preservation planning. In *Proc. of IPRES 2011*, 2011.
- [10] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*, volume 1 of *Synthesis Lectures on the Semantic Web: Theory and Technology*. Morgan & Claypool, 2011.
- [11] W. Kilbride. Preservation planning on a spin cycle. *DPC What's New*, 28, 2010.
- [12] H. Kulovits, A. Rauber, M. Brantl, A. Schoger, T. Beinert, and A. Kugler. From TIFF to JPEG2000? Preservation planning at the Bavarian State Library using a collection of digitized 16th century printings. *D-Lib*, 15(11/12), November/December 2009.
- [13] G. W. Lawrence, W. R. Kehoe, O. Y. Rieger, W. H. Walters, and A. R. Kenney. Risk management of digital information: A file format investigation. Technical report, Cornell University Library, 2000.
- [14] D. Pearson. AONS II: continuing the trend towards preservation software 'Nirvana'. In *Proc. of IPRES 2007*, 2007.
- [15] D. Tarrant, S. Hitchcock, and L. Carr. Where the semantic web and web 2.0 meet format risk management: P2 registry. *The International Journal of Digital Curation*, 1(6):165–182, June 2011.
- [16] C. Weihs and A. Rauber. Simulating the effect of preservation actions on repository evolution. In *Proc. of iPRES 2011*, pages 62–69, Singapore, 2011.

Assessing Digital Preservation Capabilities Using a Checklist Assessment Method

Gonçalo Antunes, Diogo Proença, José Barateiro, Ricardo Vieira, José Borbinha
INESC-ID Information Systems Group, Lisbon, Portugal
{goncalo.antunes, diogo.proenca, jose.barateiro, rjcv, jlb}@ist.utl.pt

Christoph Becker
Vienna University of Technology
Vienna, Austria
becker@ifs.tuwien.ac.at

ABSTRACT

Digital preservation is increasingly recognized as a need by organizations from diverse areas that have to manage information over time and make use of information systems for supporting the business. Methods for assessment of digital preservation compliance inside an organization have been introduced, such as the Trustworthy Repositories Audit & Certification: Criteria and Checklist. However, these methods are oriented towards repository-based scenarios and are not geared at assessing the real digital preservation capabilities of organizations whose information management processes are not compatible with the usage of a repository-based solution. In this paper we propose a checklist assessment method for digital preservation derived from a capability-based reference architecture for digital preservation. Based on the detailed description of digital preservation capabilities provided in the reference architecture, it becomes possible to assess concrete scenarios for the existence of capabilities using a checklist. We discuss the application of the method in two institutional scenarios dealing with the preservation of e-Science data, where clear gaps were identified concerning the logical preservation of data. The checklist assessment method proved to be a valuable tool for raising awareness of the digital preservation issues in those organizations.

Categories and Subject Descriptors

H.1 [Information Systems]: Models and Principles; J.1 Administrative Data Processing Government; K.6.4 Management of computing and Information Systems

General Terms

Management, Documentation, Measurement, Verification

Keywords

Repository Audit and Certification, Trust, Digital Preservation, Reference Architecture, Checklist Assessment

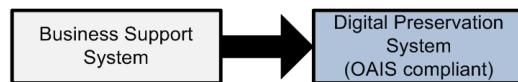
1. INTRODUCTION

Digital preservation (DP) has traditionally focused on repository-based scenarios, mainly driven by memory institutions. All the main reference models of the field such as the well-known case of the OAIS [1] have been developed with this concern in mind. These models define preservation processes, policies, requirements, and building blocks that can be used by institutions that host or want to host a repository system to effectively manage its implementation and/or its operation.

These references are widely considered valid for these kinds of scenarios. However, DP is starting to be acknowledged as a need by organizations from different walks of life in scenarios where common information systems are used for processing and managing data, and where no separate system for preservation is desirable, so that a repository approach is not applicable. These

scenarios present emergent DP requirements, where DP is seen as a *desirable property of information systems*, and not as the main source of functional requirements. In that sense, those organizations execute information management processes that cannot be aligned with the functional aspects and information structures defined in the main reference frameworks of the DP domain. Despite the apparent shift, the main objective of preservation is maintained intact, which involves assuring that information that is understood today can be transmitted into an unknown system in the future and still be correctly understood then. Thus, besides the traditional repository scenario, an alternative scenario should be considered, where DP is seen as a capability that can be added to systems. Figure 1 depicts the two possibilities.

Traditional Scenario: Digital Preservation as a System/Service



Alternative Scenario: Digital Preservation as a Capability



Figure 1. Digital Preservation Scenarios

With this in mind, a capability-based reference architecture was produced in the context of the SHAMAN¹ project and described in [3]. Reference architectures have the aim of capturing domain-specific knowledge and integrate that knowledge in a way that it can be later reused for developing new system architectures for the domain in question [4]. In that sense, the capability-based reference architecture captures knowledge from the DP domain, consolidates that knowledge taking into account reference models and best-practices of related or highly relevant domains, so that it can be reused for assessing and guiding the integration of DP capabilities in information systems. The purpose is to deliver value in organizations where DP is not a business requirement, but it required to enable the delivery of value in the primary business.

Several assessment methods are currently available in the DP domain for evaluating the effectiveness of DP in repository-based scenarios. Works like the Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) [5], DRAMBORA [6], or the freshly published ISO 16363:2012 [7], allow the assessment of a repository system and the surrounding

¹ <http://shaman-ip.eu/>

organizational environment using several different perspectives. However, their application in non-traditional DP scenarios is difficult, mainly due to the assumption that a repository system is present and that once data enters such system, it will only be accessed again in the long-term. This work proposes a checklist assessment method based on the capability-based reference architecture. The checklist itself is based on the assessment methods already existing in the DP domain, but significantly reworked and aligned with the capability approach, so that it can be applied to any scenario. It contains sets of criteria organized per capability. The implementation was made through a spreadsheet that can be configured by the user in order to concede different weights to different criteria according to the concerns of the stakeholder filling the checklist. In that way, the current DP capabilities can be identified and their levels assessed, and a gap analysis between the current and the desired situation can be performed, which can support decision making on improvements to the current situation.

This paper is structured as follows. Section 2 describes the related work in terms of assessment checklists in the DP domain and in other relevant domains. Section 3 describes a capability-based reference architecture for DP. Section 4 describes a method for assessing the current DP capabilities of an organization and a companion checklist for performing the assessment. In Section 5, the application of the checklist assessment method to two institutions dealing with the issue of preserving e-Science data is described. Finally, Section 6 discusses lessons learned, and draws conclusions.

2. RELATED WORK

The usage of assessment checklists is widely spread, being used in various areas. In the DP domain, the Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) [5] is one example. Its purpose is to be an audit and certification process for the assessment of the trustworthiness of digital repositories, and its scope of application it's the entire range of digital repositories. It is based on the OAIS model [1]. The final version of TRAC was published in 2007, based upon the experience and findings of various test audits by the Center for Research Libraries from 2005 to 2006. It contains 84 criteria which are divided into three main sections: Organizational infrastructure; Digital object Management; and Technologies, technical infrastructure, and security. Within each of this sections are various subsections and under the subsections are the criteria. A successor version of TRAC, a standard for Trusted Digital Repositories (TDR), was published by ISO in February 2012, the ISO16363:2012 standard [6].

In the DP domain there are other assessment tools, for example, the Northeast Document Conservation Center self-assessment tool [8]. This tool aims at helping the museums, libraries, archives, and other cultural organizations to begin thinking about long-term sustainability of their digital collections and complements the DP readiness assessment developed by the same center. It covers the following topics: (1) Mission and Goals; (2) Policies and procedures; (3) Staffing; (4) Finances; (5) Digital content; (6) Technology; (7) Access and metadata; (8) Digital preservation and (9) Rights Management.

A different approach for the assessment of repositories has been taken by DRAMBORA [6], a digital repository audit method based on risk assessment. DRAMBORA characterizes digital curation as a risk-management activity, because it recognizes the job of a digital curator as the rationalization of the uncertainties

and threats that inhibit efforts to maintain digital object authenticity and understandability, transforming these into manageable risks. There are six stages within the process. The first stages require that auditors develop an organizational profile, describing and documenting the repository's mandate, objectives, activities and assets. Then, risks are derived from each of these, and assessed in terms of their likelihood and potential impact. In the end, auditors are encouraged to conceive of appropriate risk management responses to the identified risk.

There are other domains which make use of checklist in order to assess a certain capability. For example in the IT domain, ISACA provides an IT Governance Self-Assessment checklist [9] in order for the management to determine, for each of the COBIT [10] processes: (1) How important they are; (2) Whether it is well performed; (3) Who performs and who is accountable; (4) Whether the process and its control is formalized and (5) Whether it is audited.

Other domains of usage include teaching [11], for example, to record observed performance of students while working in groups, to keep track of progress over time or even help students fulfill task requirements.

In conclusion, assessments using checklists are well spread in numerous domains, including the DP domain, applied for example in healthcare institutions [13], pharmaceutical industry, and manufacturing, and many other areas as described in [14] and [15]. Checklists are proven to be a successful tool to verify the state of certain aspect, in an organization, class room or even yourself.

However, DP assessment checklists assume the presence of a repository system and that once data enters the repository it will be seldom accessed. Despite that being desirable for a wide range of scenarios (e.g., cultural heritage), the existence of such solution might not be adequate for determined organizations, where data management processes are well-defined and established and specialized information systems are in place. In other words, this work aims to bridge that existing gap through a proposal of a capability assessment checklist that can be applied to any organization. Additionally, while existing DP checklists allow the assessment of important aspects of DP in organizations, they do not provide a means for evaluating the current capability level. This alone allows performing a gap analysis that can help organizations to make investments in order to fill the gaps.

3. A CAPABILITY-BASED REFERENCE ARCHITECTURE FOR DIGITAL PRESERVATION

A reference architecture can be defined as a way of documenting good architectural practices in order to address a commonly occurring problem through the consolidation of a specific body of knowledge with the purpose of making it available for future reuse [4]. Thus, a reference architecture for DP provides a way of capturing the knowledge of the DP domain, so that it can be instantiated in concrete architectures for real system implementations.

In recent years several DP reference models and frameworks have been developed providing terminology, building blocks, and other types of knowledge derived from an in-depth analysis of the domain. Although being widely accepted, these reference models are not aligned among themselves and often overlap with established references and models from other fields, such as IT Governance or Risk Management. Moreover, those frameworks

Table 1. Reference Architecture Capabilities

Capability		Description
GRC Capabilities	GC1. Governance	The ability to manage and develop the services, processes and technology solutions that realize and support DP capabilities. This includes engaging with the designated communities in order to ensure that their needs are fulfilled is also an important aspect. The ability to negotiate formal succession plans to ensure that contents do not get lost is another important aspect.
	GC2. Risk	The ability to manage and control strategic and operational risks to DP and opportunities to ensure that DP-critical operations are assured, including the sustainability of those operations and disaster recovery.
	GC3. Compliance	The ability to verify the compliance of DP operations and report deviations, if existing. Certification is also an important aspect of this capability and it consists in the ability to obtain and maintain DP certification status.
Business Capabilities	BC1. Acquire Content	The ability to offer services for transferring content from producers into the organization's systems. This includes services for reaching agreement with producers about the terms and conditions of transfer.
	BC2. Secure Bitstreams	The ability to preserve bitstreams for a specified amount of time (Bitstream preservation).
	BC3. Preserve Content	The ability to maintain content authentic and understandable to the defined user community over time and assure its provenance. (Logical preservation).
	BC4. Disseminate Content	The ability to offer services for delivering content contained in the organization's systems to the user community or another external system. This includes services for reaching agreement about the terms and conditions of transfer.
Support Capabilities	SC1. Manage Data	The ability to manage and deliver data management services, i.e. to collect, verify, organize, store and retrieve data (including metadata) needed to support the preservation business according to relevant standards.
	SC2. Manage Infrastructure	The ability to ensure continuous availability and operation of the physical, hardware, and software assets necessary to support the preservation.
	SC3. Manage HR	The ability to continuously maintain staff which is sufficient, qualified and committed to performing the tasks required by the organization.
	SC4. Manage Finances	The ability to plan, control and steer financial plans and operations of the organization's systems to ensure business continuity and sustainability.

are not always aligned with best practices, resulting in specifications that are not easy to use or that are not reusable at all. A reference architecture following best practices in the field of enterprise architecture would fit the purpose of making that knowledge available in a way that would facilitate its reuse.

In order to create a DP reference architecture that infused domain knowledge, the TOGAF Architecture Development Method (ADM) [12] was used for developing an architecture vision accommodating DP capabilities. For that, the main reference models of the domain were surveyed and integrated, providing a means of effectively addressing the issues of DP, while providing a bridge for the development of concrete DP-capable architectures. Following the ADM, the stakeholders of the domain and their concerns were identified along with the drivers and goals. This resulted in a set of general DP capabilities derived from the context, in a process that is documented in [13].

A capability is not a business function, but an ability realized by a combination of elements such as actors, business functions and business processes, and technology, and it must be related with at least one goal. This reference architecture for DP defines a set of capabilities that can be divided in three groups, which are also described in an increased level of detail in Table 1:

Governance, Risk and Compliance (GRC) Capabilities - Governance capabilities are required to manage the scope, context and compliance of the information systems in order to ensure fulfillment of the mandate, continued trust of the external stakeholders and sustainable operation of the systems.

Business Capabilities - Business capabilities are required to execute a specified course of action, to achieve specific strategic goals and objectives.

Support Capabilities - Support capabilities are required for ensuring the continuous availability and operation of the infrastructure necessary to support the organization, including physical assets, hardware, and software.

Table 2. Goals and Capabilities

ID	Goals	Capabilities
G1	Acquire Content...	BC1;
G2	Deliver...	BC4;
G3	...preserve provenance...	BC2, BC3, SC1;
G4	...preserve objects...	BC2, BC3;
G5	React to changes...	GC1, GC2, BC3, SC2;
G6	...sustainability...	GC1, GC2, GC3, SC2, SC3, SC4;
G7	Build trust...	GC1, GC2, GC3;
G8	Maximize efficiency...	GC1, GC2, SC1, SC2, SC3, SC4;

The reference architecture also defines general goals for DP. Eight goals were derived from the various references collected: (i) G1. **Acquire content** from producers in accordance to the mandate, following agreed rules; (ii) G2. **Deliver** authentic, complete, usable and understandable objects to designated user community; (iii) G3. Faithfully **preserve provenance** of all objects and deliver accurate provenance information to the users upon request; (iv) G4. Authentically **preserve objects** and their dependencies for the specified time horizon, keeping their integrity and protecting them from threats; (v) G5. **React to changes** in the environment timely

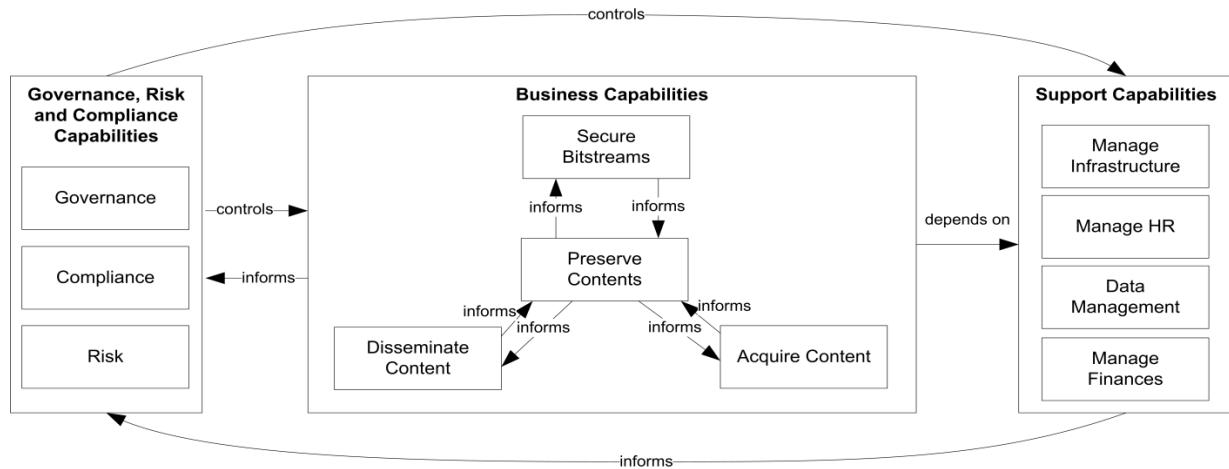


Figure 2. Capability Relationship Diagram

in order to keep objects accessible and understandable; (vi) G6. Ensure organization's **sustainability**: mandate, technical, financial, operational, communities; (vii) G7. Build **trust** in the depositors, the designated community and other stakeholders; and (viii) G8. **Maximize efficiency** in all operations. Table 2 provides a consolidated overview of all goals and the related capabilities considered here.

The categorization of these capabilities of course is partly context-dependent: in a concrete business environment, DP will generally be seen as a part of IT Governance and thus of Governance. Since it is our core focus of modeling, DP is highlighted and as such presented separately from more general aspects of IT Governance. Upon realization in a business environment, DP (and Data Management) will likely be realized as part of IT Governance, and will thus be submerged into it.

Capabilities do not exist in isolation and will have mutual dependencies. A model of their relationships and the specification of the relations existing between capabilities enable operationalization of these capabilities and an assessment of the influences exerted within capabilities in concrete scenarios. Table 3 describes the different types of relations that may exist between capabilities.

Table 3. Relations between Capabilities

Name	Description
influence	A directed relation between two capabilities
controls	An influence that determines the range of possible behavior
informs	An influence that does not exert full control, but constitutes a flow of information (that may drive or constrain the range of possible behavior in a non-exclusive way)
depends on	A relation that constitutes a dependency: The using capability is unable to act without relying on capabilities offered by the used capability. This implies a reverse "informs" relationship.

Figure 2 depicts the relations existing between capabilities. At the top level, GRC capabilities exert control over Business capabilities and Support capabilities, since they set out the scope and goals for business, and represent the regulators that constrain business. Business capabilities inform the GRC capabilities, in

particular: (i) Governance, to provide information about the operations and the status of the organization's systems, to assess opportunities and potential and be aware of operational constraints, and to determine the adequacy of means to achieve ends; (ii) Compliance, to enable auditing of compliance to regulations; and (iii) Risk, to provide information about the adequacy of preservation actions to face threats endangering the preserved contents. Support capabilities inform GRC capabilities since GRC needs information to successfully govern support capabilities. Business capabilities also have a dependency relationship with Support capabilities, since the former relies on the later. Although other relation types may exist between top-level capabilities, only the most prevalent are depicted on the diagram.

As for the relationships between Business capabilities, the Acquire Content capability informs the Preserve Contents capability, since the former constitutes a system boundary and thus the point where the organization gets control of content and the properties of acquired content are of interest for preservation. The same relationship is also true in the opposite direction since the limits of operational preservation may constrain the range of contents that can be accepted responsibly. The Disseminate Content informs the Preserve Contents since Dissemination requirements may drive and/or constrain preservation. Again, the same relationship is also true in the opposite direction since the limits of operational preservation may constrain the options for dissemination. The Secure Bitstreams capability informs the Preserve Contents capability since the way the bitstreams are physically secured may drive or constrain preservation (i.e. probabilities for bit corruption). The same relationship is also true in the opposite direction since effects of preservation may drive or constrain the way the bitstreams are physically secured (i.e. file sizes). For a detailed discussion on the existing relationships, please refer to [12].

4. ASSESSING DIGITAL PRESERVATION CAPABILITIES

With the detailed description of capabilities provided, it becomes possible to assess concrete scenarios for the existence of capabilities, since the breakdown provided allows easier assessment of the organization, making the bridge into the business architecture. An organization should map the stakeholders and their concerns in the ones provided in the reference architecture [13]. Based on that, the preservation drivers

and goals are determined, also based on the ones provided by this reference architecture, but also checking at all times for possible constraints that might affect the architecture work. That process shall provide a clear vision of the current DP capabilities and the ones effectively missing. The next following section provides a method to be used together with a checklist. After the assessment, the development and deployment of capabilities in concrete scenarios becomes possible through the development of architecture viewpoints, following the TOGAF ADM Business Architecture phase.

This section describes a checklist-based method for assessing an organization for its DP capabilities.

4.1 Checklist Assessment Method

The Checklist Assessment Method comprises five steps, as shown in Figure 3. It requires a companion checklist document, described in the following subsection. The first three phases deal with setting the organizational context. The two last steps respectively deal with the application of the checklist for determining which DP capabilities are currently deployed in the organization and their current level of effectiveness.

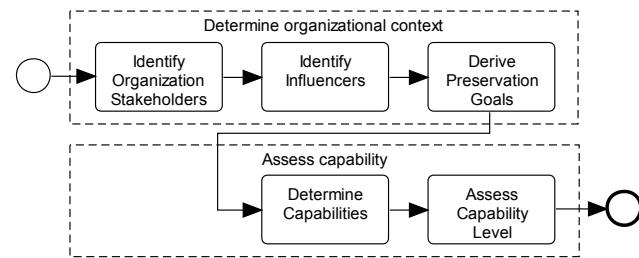


Figure 3. Checklist Assessment Method

- 1. Identify Stakeholders:** This first step deals with the identification of the stakeholders in the organization referring to the stakeholders defined in the reference architecture [13]. Since stakeholders in the organization might not be named as the ones described, they can be mapped to one or more stakeholders of the organization. For that identification, it is essential that the key questions and concerns of each stakeholder are taken into account.
- 2. Identify Influencers:** After the identification of the stakeholders, it will be possible to identify the influencers according to their concerns. For that, the list of influencers present in the reference architecture [13] should be used. Note that both drivers (which motivate the organization to set its goals) and constraints (which might constrain the deployment of means and the achievement of goals) should be identified.
- 3. Derive Preservation Goals:** The drivers derived in the previous step can then be used for deriving specific preservation goals for the organization. Those goals should be based on the generic goals provided in the reference architecture [13].
- 4. Determine Capabilities:** Then, according to the defined goals and their relationship to the capabilities, the capabilities needed to achieve the goals for the specific case should be determined, using for that purpose the checklist described in the next subsection.

- 5. Assess Capability Level:** Using the checklist, the capability level of a given organization in certain period of time can be verified. The checklist is divided into three main sections, one for each top-level capability (GRC, Business and Support). Then these sections are divided into their constituent sub-capabilities. With results given by the checklist, a gap analysis can be performed to check the current level of capability, compare it with

the organization goals or compare between different points in time.

4.2 The Assessment Checklist

Table 4 depicts an excerpt of the capability assessment checklist. The compliance criteria are based on references of the area of DP, especially on TRAC, which were reworked in order to be aligned with the capability approach followed in this work, thus loosing the repository-orientation. In other words, mentions to the concept of repository where removed and when possible, repository-specific criteria were reworked and generalized in order to widen the scope of application to all types of information systems. When the adaptation was not possible, the requirements where still accommodated in the checklist, although with a note stating the conditions to which the criteria apply.

Table 4. Excerpt of the Capability Assessment Checklist

No.	Criteria	Y/N
GC	GRC Capabilities	
GC1	Governance	
GC1.1	The organization has a documented history of the changes to its operations, procedures, software, and hardware.	
GC1.2	The organization has issued a statement that reflects its commitment to the long-term retention, management and access to digital information that is accessible to its community of users.	
GC1.3	The organization has defined the potential community(ies) of users and associated knowledge base(s) required for understanding information.	
GC1.4	The organization has publicly accessible definitions and policies in place to dictate how its preservation requirements will be met.	
GC1.5	The organization has policies and procedures to ensure that feedback from producers of information and users is sought and addressed over time.	

The idea behind the checklist is that any organization of any domain and with any type of information systems deployed can be able to apply it and check its current DP capabilities.

This checklist is available as a spreadsheet, allowing two methods for calculating the compliance level: automatic, which is a linear method; and custom in which we can define the weights for each criterion.

Each capability group is measured from 0% to 100% of compliance. Then each sub-capability has a maximum percentage which in the custom evaluation method can be defined. For instance, if we want the Governance capability (GC1) to weight 50% of the Governance Capability (GC) group, then we can add the weights 32% for the Risk capability (GC2) and 18% for the Compliance capability (GC3) (Note that the total amount for GC, GC1+GC2+GC3, has to be 100%). If we want to define custom weights for the GC1 criteria, for example, GC1 has a maximum of 50%, so we want GC1.1 to weight 5%, GC1.2 to weight 15%, GC1.3 to weight 10%, GC1.4 15%, GC5 5% and the others 0%. Finally, we want GC2 and GC3 to be calculated evenly between the criteria. Figure 4 depicts the customization of GC1. The compliance levels can also be adapted using the table pictured in Table 5.

In order produce a gap analysis with the results achieved, the organization's compliance level target for each capability must be

provided in the ‘questionnaire’ spreadsheet, as an organization might set its own goals concerning the deployment of capabilities due to a variety of reasons (e.g., cost, schedule, etc.) This is pictured in Figure 5.

Capabilities	Weights
GC	50
GC1	50
GC1.1	5
GC1.2	15
GC1.3	10
GC1.4	15
GC1.5	5
GC1.6	0
GC1.7	0

Figure 4. Assigning Weights to Capabilities

Table 5. Compliance Levels Configuration

Levels		
Levels	Percentage	
	Min.	Max.
1	0	25
2	26	45
3	46	65
4	66	80
5	81	100

Percentage	Level	Target	Difference
0,00	1	5	-4
0,00	1	4	-3

Figure 5. Gap Analysis Configuration

After filling the questionnaire, results can be observed by the means of spider graphs. Figure 6 depicts the compliance levels of a fictional company, the organization XYZ, determined using the companion checklist. In the top-left we can see the global compliance level regarding the three main capabilities depicted in this document. The additional graphs depict the compliance levels for each of the three top-level capabilities. There are three lines in each of these figures: one for organization’s target which is the compliance level that the organization wants to achieve, another line for the first compliance check (start) which is the result achieved by the organization on the first compliance check, and finally, another line for the actual compliance level which should be refreshed through time in each compliance check. The main goal here is for the stakeholders to check periodically if their concerns are being correctly addressed through time.

5. ASSESSMENT APPLIED TO TWO E-SCIENCE INSTITUTIONS

e-Science concerns the set of techniques, services, personnel and organizations involved in collaborative and networked science. It includes technology but also human social structures and new large scale processes of making science.

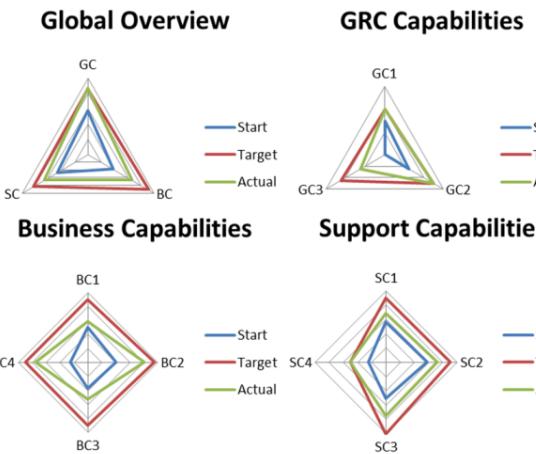


Figure 6. Compliance Graphs

DP is recognised as a required property for future science, to assure communication over time, so that scientific information that is understood today can be reused in the future to produce new knowledge [16].

To conduct a systematic assessment of the preservation capabilities of research organizations, the checklist assessment method was applied to two selected institutions with preservation scenarios dealing with e-Science data: a Civil Engineering (structure monitoring data) and high-energy physics (experimental data). A meeting was scheduled with both groups of stakeholders in which the issues surrounding DP in e-Science scenarios were described along with the reference architecture. After that, the stakeholders were asked to apply the checklist assessment method.

5.1 High Energy Physics Institution

The institution approached is responsible for several studies in the fields of high energy physics (HEP). It is also engaged in large scale experiments occurring in the context of international collaborations. Due to the special characteristics of each of those experiments and the associated costs, it is highly unlikely that the data obtained in that context can be fully reproduced in a new experiment. This fact presents a motivation for preserving this data, since with the development of new theoretical models it might highly relevant to perform a reanalysis of the produced data. The preservation of this data is a challenging task due to the fact that researchers of participating institutions perform local analysis of that data, using for that purpose specific data software which might make use of custom modules developed by the researcher himself, data analysis tools, simulation software, and other scripts. Each of the steps in the analysis might produce different types of intermediate data, each one stored in a determined format.

Table 6 depicts an excerpt of the checklist that was filled by a HEP stakeholder for the Risk capability. The “x” indicates that the criterion is being fulfilled, and the “0” indicates otherwise. We see that two criteria are not met by the organization.

The overall results of the assessment for the high energy physics scenario can be observed in Figure 7. Since this is in fact a first assessment, only the Start and Target lines are displayed. The global overview indicates that Support capabilities are at the level 4 out of 5 of compliance, while Governance and Business capabilities are at level 2 out of 5 of compliance. Through the observation of the GRC capabilities graph, it is possible to see that

the governance and compliance capabilities are at a very low level. The Business capability graph indicates that the Preserve Contents capability is almost non-existent, while the Secure Bitstreams capability is at the level 4 out of 5. Finally the Support capabilities graph shows that the Manage Data and the Manage HR capabilities need improvement.

Table 6. Risk Capability Assessment for the HEP Institution

GC2	Risk	
GC2.1	The organization has ongoing commitment to analyze and report on risk and benefit (including assets, licenses, and liabilities).	x
GC2.2	The organization has a documented change management process that identifies changes to critical processes that potentially affect the organization and manages the underlying risk.	0
GC2.3	The organization has a process for testing and managing the risk of critical changes to the system.	x
GC2.4	The organization has a process to react to the availability of new software security updates based on a risk-benefit assessment.	x
GC2.5	The organization maintains a systematic analysis of such factors as data, systems, personnel, physical plant, and security needs.	x
GC2.6	The organization has implemented controls to adequately address each of the defined security needs.	x
GC2.7	The organization has suitable written disaster preparedness and recovery plan(s), including at least one off-site backup of all preserved information together with an off-site copy of the recovery plan(s).	0

It is possible to conclude from the analysis that the knowledge about the implications of DP was somewhat lacking: The organization has a strong capability level for securing bitstreams, the capability of performing the logical preservation of objects is at a very low-level. This is also noticeable in the fact that the capabilities concerning the governance and compliance of preservation are also very low, which indicates that top-level management is not aware of the need to perform effective preservation of the scientific production.

5.2 Civil Engineering Institution

The civil engineering institution approached is responsible for the monitoring of large civil engineering structures to ensure their structural safety, which is achieved through the usage of automatic and manual data acquisition means for real-time monitoring and automatically trigger alarms, when needed. The collected data is then transformed and stored in an information system where it can be later accessed and analyzed. The motivation for preserving this data comes from different aspects such as the fact that it is unique and cannot be produced again, legal and contractual compliance issues are involved, and that its future reuse is highly desirable since new research on the behavior of structures can be performed. The preservation of this data raises several challenges due to the fact that a large variety of sensors are used, making use of different representations for organizing data, and that a large variety of data transformation algorithms can be applied to data.

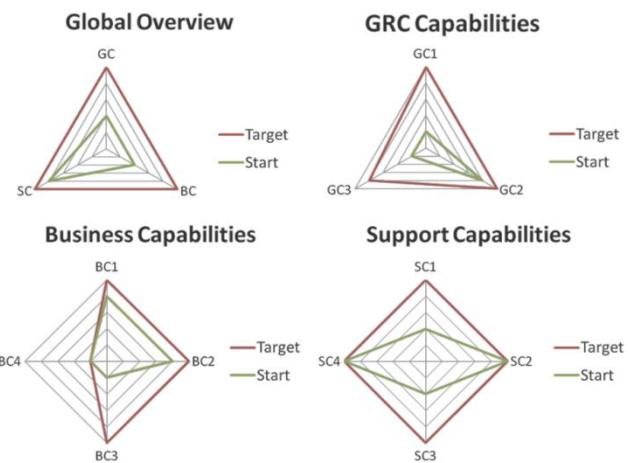


Figure 7. Compliance Assessment for the High Energy Physics Institutions

Table 7. Secure Bitstreams capability assessment for the civil engineering institution

BC2	Secure Bitstreams	
BC2.1	The organization provides an independent mechanism for audit of the integrity of all the data.	x
BC2.2	The organization implements/responds to strategies for the secure storage of objects and storage media migration in order to perform bitstream preservation of digital objects.	x
BC2.3	The organization actively monitors integrity of digital objects.	x
BC2.4	The organization reports to its administration all incidents of data corruption or loss, and steps taken to repair/replace corrupt or lost data.	x
BC2.5	The organization has effective mechanisms to detect bit corruption or loss.	0

Only operational stakeholders were available for applying the checklist assessment, which limited the assessment to the business capabilities. Table 7 depicts an excerpt of the checklist filled by a civil engineering stakeholder for the Secure Bitstreams capability. Only one of the criterions was not being filled.

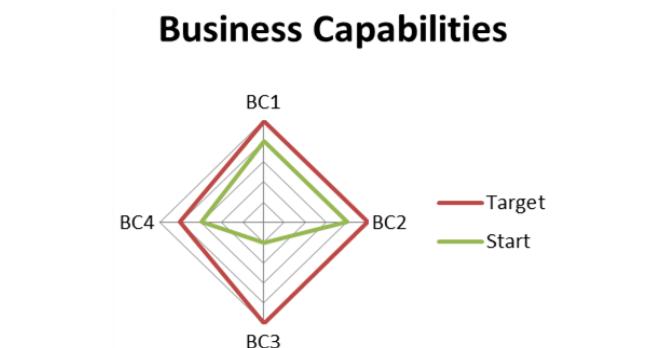


Figure 8. Assessment of Business Capabilities in the Civil Engineering Scenario

Figure 8 depicts the results of the assessment of business capabilities. The assessment determined that the Preserve

Contents capability is almost non-existent, while the Disseminate Content capability needs improvement. From the analysis of the results, it can be concluded again that the knowledge about what sets DP apart from bitstream preservation is very low, since despite having high bitstream preservation capabilities, the capabilities concerning logical preservation are very low. This might be the potential reason for also having low content dissemination capabilities.

6. CONCLUSIONS AND OUTLOOK

This article presented and evaluated a checklist-based method for capability assessment in digital preservation. The method presented is based on a capability-based reference architecture for DP that aims to provide guidance in the assessment and integration of DP capabilities into the information systems of organizations. For that purpose a checklist aimed to be used together with the method was described. The checklist provides sets of criteria for each DP capability which then can be used for evaluating the current level of the DP capabilities of an organization and the gap between current and desired capability, and in that way determining which strategic options can be taken in order to improve capability levels. It was implemented in a way that it can be configured by the stakeholders, allowing changing the weights of the criteria according to the concerns of the stakeholders of the organization being assessed.

The implemented checklist was then applied to two institutions dealing with the need for preserve e-Science data: a High Energy Physics institution and a Civil Engineering institution. From the results of the application, we can conclude that the knowledge of the implications of the logical preservation of data is not well known, despite the existence of bitstream preservation capabilities. This is a commonly observed phenomenon, since many organizations are moving step-by-step from physically securing bitstreams to ensuring continued access to the encoded information. The state of capabilities is also reflected on the level of the governance and compliance capabilities which indicates that the issue is mainly seen as a technological issue, disregarding all the policy aspects that are so important to DP.

The application of the checklist to the two institutions was considered valuable by the involved stakeholders, as it raised awareness of the different aspects involved in the preservation of data. Additionally, the resulting assessment provided an overall picture of the current DP capabilities. Nonetheless, despite providing hints about the possible solutions to the identified gaps, the assessment does not provide concrete and clear answers in terms of solutions to the identified issues. Due to recognizing that need, current and future work focuses on the development of techniques for the modeling and visualization of DP capability patterns so that capabilities can be designed and implemented based on a capability pattern catalog after an assessment has been performed.

7. ACKNOWLEDGMENTS

This work was supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds and by the projects SHAMAN and TIMBUS, partially funded by the EU under the FP7 contracts 216736 and 269940.

8. REFERENCES

- [1] ISO 14721:2010. Space data and information transfer systems – Open archival information system – Reference model. 2010.
- [2] Becker, C., Antunes, G., Barateiro J., Vieira R. 2011. A Capability Model for Digital Preservation. In *Proceedings of the iPRES 2011 8th International Conference on Preservation of Digital Objects*. (Singapore, November 01 – 04, 2011).
- [3] Cloutier, R., Muller, G., Verma, D., Nilchiani, R., Hole, E., Bone, M. *The Concept of Reference Architectures*. Wiley Periodicals, Systems Engineering 13, 1 (2010), 14-27.
- [4] RLG-NARA Digital Repository Certification Task Force. 2007. *Trustworthy repositories audit and certification: Criteria and checklist*. OCLC and CRL, http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf (accessed 18 May 2012).
- [5] McHugh, A., Ruusalepp, R. Ross, S. & Hofman, H. *The Digital Repository Audit Method Based on Risk Assessment*. DCC and DPE, Edinburgh. 2007.
- [6] ISO 16363:2012. Space data and information transfer systems – Audit and certification of trustworthy digital repositories. 2012.
- [7] Bishhoff, L., Rhodes, E. 2007. Planning for Digital Preservation: A Self-Assessment Tool. Northeast Document Conservation Center, <http://nedcc.org/resources/digital/downloads/DigitalPreservationSelfAssessmentfinal.pdf> (accessed 18 May 2012).
- [8] Mansour, C. 2008. *IT Governance and COBIT*. Presentation. ISACA North of England Chapter, <http://www.isaca.org.uk/northern/Docs/Charles%20Mansour%20Presentation.pdf>, Leeds. 2008.
- [9] IT Governance Institute. *COBIT 5 – A business Framework for the Governance and Management of Enterprise IT*. 2012.
- [10] Center for Advanced Research on Language Acquisition. Evaluation – Process: Checklists. University on Minnesota, http://www.carla.umn.edu/assessment/vac/evaluation/p_3.html, Minneapolis, MN. 2012.
- [11] The Open Group. *TOGAF Version 9*. Van Haren Publishing. 2009.
- [12] SHAMAN Consortium. *SHAMAN Reference Architecture v3.0*. http://shaman-ip.eu/sites/default/files/SHAMAN-REFERENCE%20ARCHITECTURE-Final%20Version_0.pdf. 2012.
- [13] Bote J., Termens M., Gelabert G. *Evaluation of Healthcare Institutions for Long-Term Preservation of Electronic Health Records*. Springer-Verlag, Centeris 2011, Part III, CCIS 221, 136-145. 2011.
- [14] Ashley L., Dollar C. *A Digital Preservation Maturity Model in Action*. Presentation. PASIG, Austin, TX. 2012.
- [15] Rogers B. *ISO Audit and Certification of Trustworthy Digital Repositories, Part II - IT Assessment Methodologies*. Presentation. PASIG, Austin, TX. 2012
- [16] *Data's Shameful Regret* [Editorial]. Nature, 461, 145. 2009.

Evaluating Assisted Emulation for Legacy Executables

Swetha Toshniwal

Geoffrey Brown

Kevin Cornelius

Indiana University School of
Informatics and Computing

Gavin Whelan

Enrique Areyan

ABSTRACT

Access to many born-digital materials can only be accomplished economically through the use of emulation where contemporaneous software is executed on an emulated machine. For example, many thousands of CD-ROMs have been published containing proprietary software that cannot be reasonably recreated. While emulation is proven technology and is widely used to run both current and obsolete versions of Windows and Unix operating systems, it suffers a fatal flaw as a preservation strategy by requiring future users to be facile with today's operating systems and software.

We have previously advocated "assisted emulation" as a strategy to alleviate this shortcoming. With assisted emulation, a preserved object is stored along with scripts designed to control a legacy software environment and access to the object enabled through a "helper" application. In this paper we significantly extend this work by examining, for a large data set, both the cost of creating such scripts and the common problems that these scripts must resolve.

1. INTRODUCTION

This paper describes a project to develop practical techniques for ensuring long-term access to CD-ROM materials. The underlying technology for our work consists of off-the-shelf emulators (virtualization software) supported by custom automation software. We use automation to capture the technical knowledge necessary to install and perform common actions with legacy software in emulation environments and hence mitigate a fundamental flaw with emulation. This work directly addresses issues of sharing through networked access to emulators and object-specific configuration and assistance.

Over the past 20 years CD-ROMs were a major distribution mechanism for scientific, economic, social, and environmental data as well as for educational materials. Our work has primarily focused upon the nearly 5,000 titles distributed by the United States Government Printing Office (GPO) under the Federal Depository Loan Program and thousands more distributed by international agencies such as UNESCO. Recently, we have expanded our study to the thousands of commercial titles held by the Indiana University Libraries. In the short-term these materials suffer from physical degradation which will ultimately make them unreadable and, in the long-term, from technological obsolescence which will make their contents unusable. Many such titles (as much as 25% of the GPO titles and perhaps more for commercial titles) require execution of proprietary binaries that depend upon obsolete operating systems and

hardware. A widely discussed technical strategy is to utilize emulation (virtualization) software to replace obsolete hardware. [6, 2, 4, 11, 10, 7, 3, 5] Recent surveys of issues related to the use of emulation in preservation based upon lessons from the Planets project include [9, 12].

A fundamental flaw with this approach is that future users are unlikely to be familiar with legacy software environments and will find such software increasingly difficult to use. Furthermore, the user communities of many such materials are sparse and distributed thus any necessary technical knowledge is unlikely to be available to library users. The work described in this paper is aimed at alleviating these issues.

As mentioned in the abstract, we have previously proposed a strategy of "assisted emulation" which attempts, through the use of helper applications and scripting, to simplify access to legacy materials. [13] In prior work we described a simple pilot study aimed at determining the basic viability of this approach. In this paper we significantly expand this work with an emphasis upon understanding the issues and difficulty associated with creating the scripts required by our strategy. In particular, we describe the results of a study involving several hundred CD-ROMs, both government and commercial through which we are able to make recommendations about the basic emulation environment, additional software requirements, and a collection of techniques used to automate access to the individual titles.

2. REVIEW OF ASSISTED EMULATION

Our approach, as described previously in [13], relies upon storing scripts along with legacy materials which are executed automatically by a "helper program" when a user accesses the materials through an ordinary web browser. For a given digital object, a single "click" causes the associated script(s) to be downloaded and executed to start and configure an emulation environment on the user's workstation. This approach is illustrated in Figure 1. Where a user requests access to an object through a browser on the client machine to a web server (1). The web server responds with a signed applet (2) causing a helper program on the client machine to execute a local emulator (3). This emulator is configured using scripts stored on the network to execute software and access objects also stored on the network. This model makes certain basic assumptions which we elaborate upon in the sequel. First, the emulator and its basic operating environment are accessible from the client machine; and second, the preserved object and any necessary scripts are accessible on networked storage.

Throughout our work we have assumed that, for a given

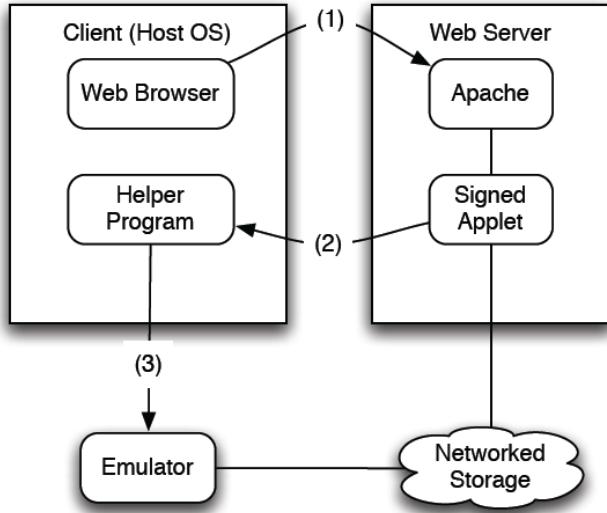


Figure 1: Client Request for Networked Resource

platform (e.g. PC, or “classic Macintosh”), most software can be accommodated by a small number of configurations. For example, when working with legacy digital objects originally supported by various versions of Windows, we have found that Windows XP provides a high degree of backwards compatibility and hence a large fraction of digital objects from prior releases of Windows can be accessed by an emulator running Windows XP. Thus, we assume that the client machine has access to an emulator with a small number of reference configurations. In this paper we concentrate upon preserving objects designed to execute under Windows and MS-DOS – in separate work we have investigated the use of emulation to preserve classic Macintosh applications. [1]

Our “reference” configuration consists of Windows XP and Adobe Reader, with other software (e.g. Office and QuickTime) installed as needed by the helper scripts. It is not feasible to combine all additional software in a single reference image because some CD-ROMs depend upon specific software versions; however, the results we present suggest that a limited set of configurations could be created to minimize the frequency with which run-time installation of helper applications is required.

As mentioned above, we assume that the digital objects and necessary scripts are accessible through networked storage (in our work we have used the Andrew File System (AFS)). [8] The objects we work with are the thousands of CD-ROMs in the Indiana University Libraries. Our archive organization is illustrated in Figure 2 which shows how uniquely numbered CD-ROM images (ISO files) are stored in eponymous directories along with item specific scripts (e.g. install.exe), and generated ISO files containing required helper software. The CD-ROM images are created with standard software from the physical disks and the scripts are created using a process described in Section 4.2. This figure differs from our previous work with the inclusion of additional ISO files to provide helper applications required by specific CD-ROM images.

3. RESEARCH ENVIRONMENT

Over the past five years we have built a research collection of nearly 5000 CD-ROM images from materials in the Indiana University Libraries. These include United States Government documents, publications of international organizations (e.g UNESCO) and foreign government, commercial publications, and educational publications. In our initial work on assisted emulation we focused upon the US Government documents which generally have minimal requirements for additional software and offered limited variety in terms of installation processes. We have greatly expanded this work and in this paper provide results based upon analyzing 1325 CD-ROMs of all types.

For our emulation platform we use VMWare Server running on top of Windows 7 (64-bit) (we also run on top of Linux in our development environment). There are many alternative emulation platforms for executing Windows operating systems with appropriate APIs enabling automation. Within the virtual machine, we run Windows XP Professional with a basic set of application programs.

Assisted emulation depends upon the creation of scripts that can be executed when a patron requests access to a particular CD-ROM image. For Windows emulation we use the freeware tool AutoIt¹, a BASIC-like scripting language that facilitates automating user tasks by capturing keystrokes, mouse movements, and window controls. While we installed AutoIt on our baseline machine, it is only required for the creation of scripts which can be conveniently converted into executable files. This is discussed further in the next section.

4. CD-ROM AUTOMATION

Our automation work consists of two phases for each CD-ROM image. In the first phase we explore the image by mounting it in our reference virtual machine to gather some

¹ AutoIt Automation and Scripting Language. <http://www.autoitscript.com/site/autoit/>

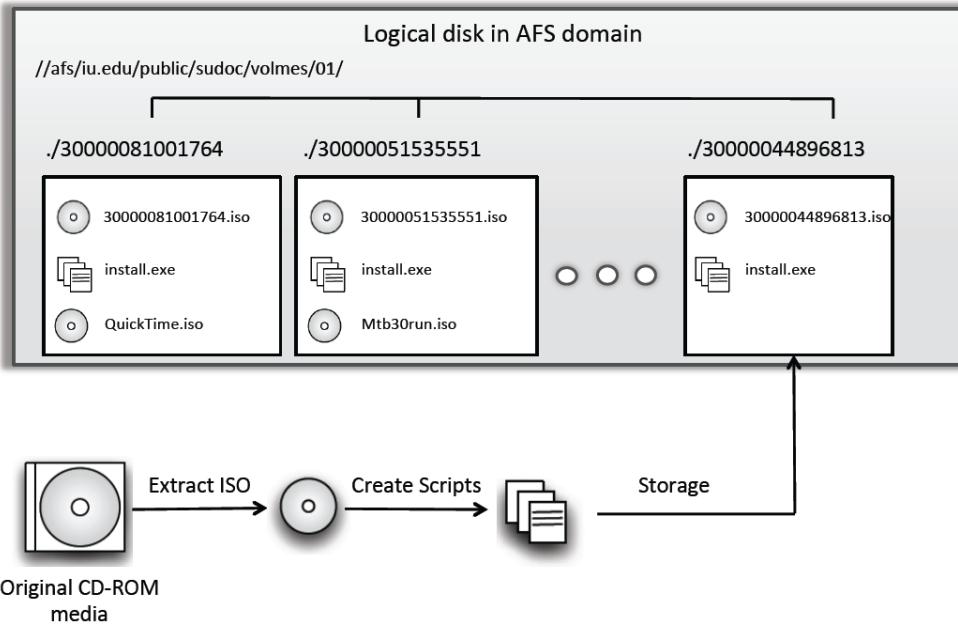


Figure 2: Organization of the Virtual Archive

basic information including: whether the image includes executable software, any required installation procedures, dependencies upon other CD-ROM images (in some cases a set of CD-ROMs are interdependent), and whether there appear to be additional software requirements. Once the requirements of a CD-ROM image are understood, the second phase consists of automating any installation processes and creating ISO files containing external software that may be required – as our repertoire has grown, we have found there is a considerable amount of reuse both in terms of scripting and these additional ISO files.

4.1 Exploration Phase

The exploration process can be quite simple – for example, many CD-ROMs contain README files that explain installation procedures and define externally required software. However, in some cases there is little or no guidance provided. Furthermore, many of the CD-ROMs we considered were in foreign languages which our research team could not readily read. Where external software is required, it is frequently difficult to determine which versions are compatible both with the CD-ROM software and our execution environment (e.g. some CD-ROMs require 16-bit QuickTime and only the last 16-bit version works correctly on Windows XP). Once the necessary software version is determined, it can be a challenge to find a copy (e.g. Multimedia Toolbox 3.0).

One of the more vexing problems we ran into was dealing with foreign languages – especially Asian. There are two aspects to this problem – our inability to read the language and the need for foreign language support in Windows. Resolving this problem typically required: (1) determining the appropriate language, (2) for east Asian languages installing Windows support, (3) configuring the appropriate language option in Windows.

We found it most efficient to install language support as part of our base image meaning that only steps (1) and (3) are necessary on a per-CD-ROM basis. In order to execute some programs provided on images it was necessary to configure various compatibility modes in Windows XP. These include changes to virtual memory settings, changing to 16-bit color, and setting file properties for specific compatibility modes. For programs designed to run in DOS mode, some images also required configuration of extended memory (XMS).

An additional complication was dealing with objects which were published on multiple CD-ROMs where there are cross-disk dependencies. For example, a program on one image might require access to a file on another image. Our current strategy is to simultaneously mount all dependent disks. This has a known limitation – VMware can only support up to three simultaneous virtual CD-ROMs. Ultimately, we may need to develop a more sophisticated helper program which will help the user to selectively mount CD-ROM images from a given set.

In summary, exploring the CD-ROM images revealed the program requirements, special cases, and required the development of strategies to handle these special cases. However, these problems are not unique to assisted emulation – even if patrons were provided access to the original CD-ROMs and machines capable of executing them, somebody would have to understand these obsolete environments sufficiently to overcome any obstacles. With assisted emulation there is at least the possibility to capture the required knowledge in scripts.

4.2 Helper Scripts

As mentioned previously, we use AutoIt for our script development. AutoIt executes simple programs in a BASIC-like language. Furthermore, the program provides a tool to

```

Run("D:\SETUP.EXE")
WinWait("Setup")
ControlClick("Setup", "", "Button1")
WinWait("", "successfully installed")
ControlClick("", "successfully installed", "Button1", "", 2)
WinWait("CD-ROM Delos")
ControlListView("CD-ROM Delos", "", "SysListView321", "Select",
ControlListView("CD-ROM Delos", "", "SysListView321",
"FindItem", "Delos"))
ControlSend("CD-ROM Delos", "", "SysListView321", "!{ENTER}")
WinWait("Delos Properties", "Shortcut")
WinClose("CD-ROM Delos")
ControlCommand("Delos Properties", "Shortcut", "SysTabControl321", "TabRight")
WinWait("Delos Properties", "Compatibility")
SendKeepActive("Delos Properties", "Compatibility")
Send("{TAB}{SPACE}")
ControlClick("Delos Properties", "Compatibility", "Button11")
ControlClick("Delos Properties", "Compatibility", "Button10")
Run("D:\WIN\DELOS.EXE")

```

Figure 3: Example Script

convert these programs into small executables (.exe files). In general, most of the complexity of these scripts comes from handling special cases – for example, setting compatibility mode for older Windows programs. Consider the script in Figure 3. The basic setup is accomplished within the first 6 lines. The remainder of the script is concerned with setting compatibility mode for the installed executable, and then running that executable.

Through the exploration of 1325 disks, we have built a script catalog to deal with the commonly occurring special cases. Some more difficult cases include autorun CD-ROMs where the `autorun.inf` fails under Windows XP, handling international CD-ROMs where window names and commands are in non-English Unicode, and installations where restarting the virtual machine is required after the software is installed. With experience, we have developed techniques to deal with these and other challenging cases. In general, we have been able to reuse script fragments to deal with many of the issues that arise in practice.

4.3 Analysis

In selecting test cases for this work, we have attempted to choose materials from a wide range of genres, languages and publication dates (1990-2010) and have analyzed 1325 CD-ROM titles. To understand the breadth of these choices, consider the Table 1 which provides a sampling of the range of these characteristics.² Any such characterization is, by necessity, somewhat arbitrary. We distinguish between commercial publications and government publications because our experience with the government printing office materials suggests that many government publications are primarily data in a limited set of formats; although, some earlier publications required installation of proprietary programs. Our selection of “genre” categories is intended to illustrate the breadth of materials – i.e. these are not all data sets. The language category is the least ambiguous. Note the relatively high fraction of Asia languages; this isn’t too surprising given the source of the materials – a major research

library. However, it also illustrates a challenging problem for archivists of such materials as installation of these disks presents a potential language barrier.

These various categories of works have widely varying software requirements. As mentioned previously identifying and finding additional software has been a major issue for this project – this is discussed further in the sequel. The work described in this paper has had its share of failures – CD-ROMs which we have not yet succeeded in executing. Finally, a key objective of this project has been to evaluate the cost of employing the strategy we are advocating. We present preliminary results on these costs.

4.4 Additional Software

In some cases, additional software requirements could be determined by the file types present on a CD-ROM, in other cases, error messages received when attempting to execute a CD-ROM provided helpful hints. It is not always necessary to find the exact version of software requested by CD-ROM documentation. For example, we found Adobe Reader to have excellent backwards compatibility – so good that we have added Adobe Reader X to our base configuration. In the few cases where images required an earlier version, our scripts uninstall Adobe Reader X and then install the required version. In other cases, the installation process requires specific versions (e.g. Office 97 or Office 2000) where one would expect better backwards compatibility. QuickTime has extremely poor backwards compatibility and it is essential that the correct version be selected. Finally, we sometimes substitute alternative packages; for example, we used Adobe Reader in place of Abapi reader (a Chinese “clone” of Adobe Reader). The Table 2 summarizes the software we installed based upon the CD-ROM requirements. The percentage of CD-ROMS requiring each software product is also provided. Unfortunately, determining acceptable version requirements for additional software is a trial and error process. Of the 1325 scripts we consider in this article, 1194 required the installation of some additional software; however, the majority can be satisfied by an enhanced baseline image including Adobe Reader, Internet Explorer, and Microsoft Office.

²This table is based upon an initial analysis of 240 CD-ROMs.

Category	Genre	Language
Commercial	Periodical/Serial	33.0%
Government	Informational	14.5%
	Historical	6.8%
	Database	5.6%
	Educational	5.6%
	Media	5.2%
	Cultural	4.4%
	Bibliography	4.0%
	Entertainment	3.6%
	Geographic	3.6%
	Geological	2.8%
	Academic	2.4%
	Survey	1.6%
	Literature	1.6%
	Biographical	1.2%
	Agricultural	0.4%
	Political	0.4%
	Statistical	0.4%
	English	49.2%
	Japanese	21.3%
	Chinese (PRC)	14.6%
	Chinese (Taiwan)	7.0%
	German	5.5%
	French	4.3%
	Hungarian	4.3%
	Czech	3.5%
	Romanian	3.5%
	Bulgarian	2.7%
	Estonian	2.7%
	Greek	2.7%
	Italian	2.7%
	Polish	2.7%
	Slovanian	2.7%
	Korean	0.8%
	Russian	0.8%
	Spanish	0.3%

Table 1: Characteristics of CD-ROMs

Program	Number	Percent
Adobe Reader	695	58%
Internet Explorer	255	21%
QuickTime	108	9%
Microsoft Office	65	5%
Windows Media Player	49	4%
Java	13	1%
Photoshop Pro	6	< 1%
Real Audio	2	< 1%
Multimedia Toolbox	1	< 1%

Table 2: Additional Software

Based upon these results, we recommend a virtual machine configuration supporting three CD-ROM drives (the limit for IDE), Adobe Reader X, and the Windows International Language Package. In most cases, Microsoft Office should be installed, but as mentioned above, there are situations where the latest version is incompatible. Thus, it may make sense to support a CD-ROM collection with a small number of base images (e.g. both with and without Office) in order to simplify run-time customization.

4.5 Failures

We have not yet succeeded in executing all of the CD-ROM images. A small number had physical errors introduced during the imaging process – this can be addressed by creating fresh images. More challenging are several CD-ROMs created for early versions of Windows (e.g. 3.1) which we have not been able to execute on either Windows XP or Windows 98. We have not yet attempted setting up a Windows 3.1 environment to test them. Unfortunately, the scripting tool we use is not available for Windows 3.1 so it is unlikely that we will achieve full automation for these cases. However, the virtual image requirements are likely to be quite small for Windows 3.1 and this may be a case where storing a custom virtual machine image along with the CD-ROM image makes sense.

4.6 Temporal Costs of Scripting

In this section we consider the per-item costs associated with setting up an assisted emulation environment. Factors setting up VMware and creating base images, which are one-time costs, are not considered. Throughout this work we have monitored the time used to create each script. As might be expected, as our team became familiar both with the scripting tool and solved some of the common automation problems, times have declined significantly. The time taken to write a script has ranged from a few minutes to 3 hours with an average of 15 minutes. The data for 1325 scripts are provided in Figure 4. Indeed, virtually any script that took longer than average meant that we encountered some challenge, often for the first time. Examples of the problems that we had to overcome include: changes to environment settings, finding required external software, language issues including support and documentation, installation of multiple programs, installation requiring multiple OS restarts, cryptic error messages, unusually lengthy and complex installations.

Notice that most of these issues are fundamental – they are due to issues with the underlying software and are not due to the scripting process. Some of these issues resulted in specific actions in the scripts (duplicating actions required for a normal installation). Among the more common cases were: language change (12%), Computer restart³ (13%),

³Language changes also require a restart

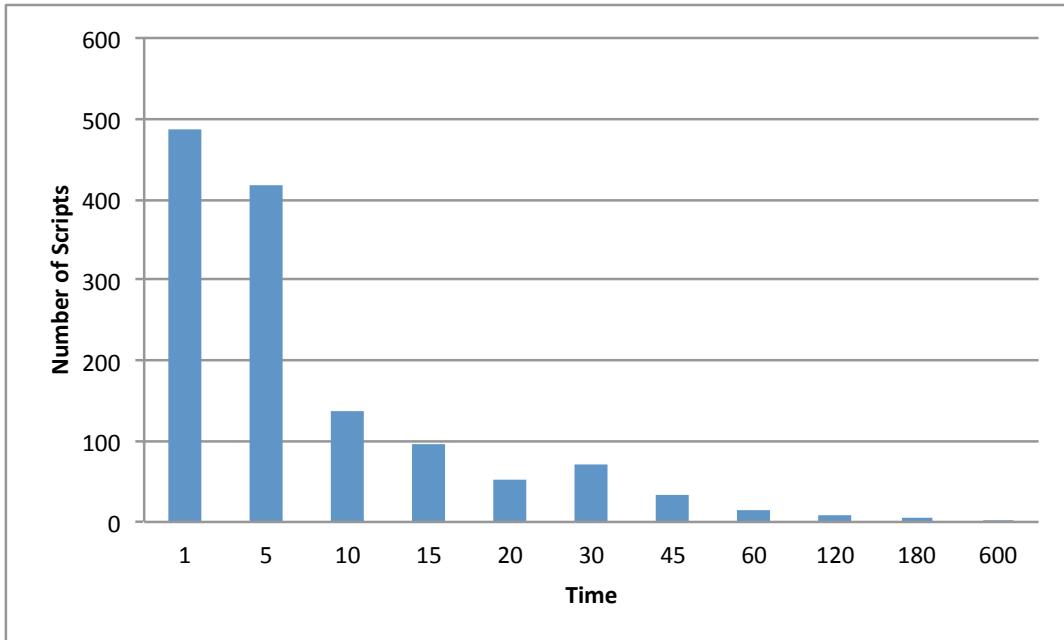


Figure 4: Number of Scripts by Creation Time (minutes)

Java installation (1.5%), free virtual memory (1.1%), display settings change (0.1%), Compatibility Settings (0.4%), and XMS memory configuration (0.1%).

4.7 Script Complexity

Another way to measure the difficulty of scripting is to consider the length of scripts. In Figure 5 we provide a chart of script lengths. The shortest scripts were a single line, for example:

```
Run("D:\start.exe")
```

which hardly justifies a dedicated script ! A more typical script, such as that illustrated in Figure 3 requires some interaction with the CD-ROM installer as well as initialization of an environment for the end-user. This example is 21 lines whereas our average script was 27.5 lines. Many of the longest scripts involved either rebooting the virtual machine during installation, changing the platform language (e.g. to support Asian languages) or installing multiple additional software applications. For example, the 158 scripts that performed language changes averaged 52 code lines. An additional 14 scripts required rebooting and averaged 68 code lines. The longest scripts which did not involve a reboot, also altered system properties (e.g. colors) to create a compatible environment for software designed to execute on older platforms.

As mentioned previously, many of these installation “tricks” are reusable – indeed they are recorded in our scripts. Consider, as an example, a fragment of a script that reboots the virtual machine during installation as illustrated in Figure 6. The key idea is that there are two phases – “prereboot” and “postreboot”. The first phase performs the basic installation (mostly elided) and, through the “_RunOnce” procedure marks a suitable variable in the registry. The postreboot procedure starts the installed application.

```
If not FileExists (...) Then
    _prereboot()
Else
    _postreboot()
EndIf

func _prereboot()
    Run('D:/SETUP.EXE')
    ...
    _RunOnce()
    Shutdown(2)
EndFunc

Func _RunOnce()
    ...
    If @Compiled Then
        RegWrite(...)
    Else
        RegWrite(...)
    EndIf
EndFunc

func _postreboot()
    ...
EndFunc
```

Figure 6: Script Performing Reboot

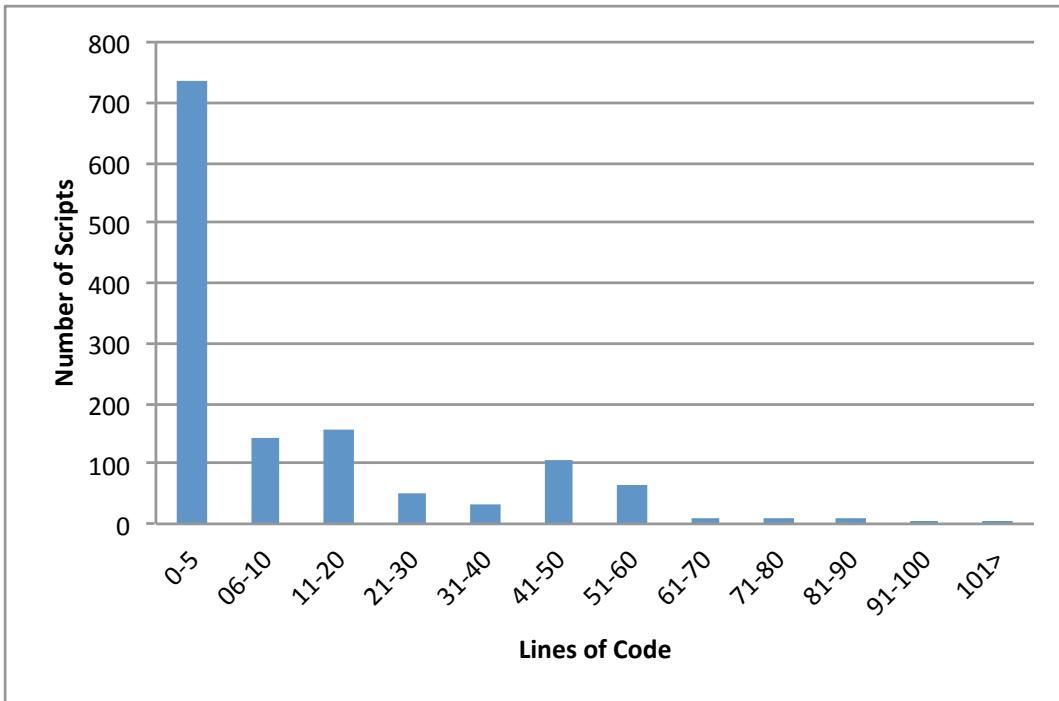


Figure 5: Lines of Code for Scripts

5. DISCUSSION

Clearly the creation of scripts is an extra layer of work required beyond installing and configuring software. The alternative would be to store a separate virtual machine image for each preserved object. For Windows XP these images are 4-8GB which implies a substantial overhead for a 500MB CD-ROM. In contrast with the development of install programs for arbitrary machines with arbitrary software configurations as is required for the development of commercial software, our scripts are required to work only in the tightly controlled environment of a virtual machine image. Furthermore, we have not found the temporal cost of writing scripts is a large additional burden. In a separate project we studied the emulation of CD-ROMs published for “classic Macintosh” machines. In that case, storing customized virtual machine images imposes a much smaller overhead (these images are typically 128MB). [1]

For many in the preservation community, the fundamental questions are how expensive is this approach and what skills are required. Most of the script development was performed by Computer Science undergraduates working as research assistants. These are bright students with some degree of programming sophistication. The data we have presented suggest that, on a per-item basis, an average of 15 minutes is required. In a more realistic production environment with the overhead of developing proper documentation and additional testing, it is reasonable to budget an hour per-item. The actual time requirements of creating the images is quite small (less than 10 minutes per item).

A side benefit of this project is that the process of creating scripts has helped us understand and collate both the common installation problems and the additional software required to preserve CD-ROM materials. In this sense, the creation of install scripts represents only an incremental ef-

fort over any principled preservation strategy.

We assumed from previous work that Windows XP would be an adequate platform for emulation of most CD-ROMs created for Windows and MS-DOS operating systems. This has proven to be largely correct; however, as we have noted, we have encountered a handful of CD-ROMs that are so tightly tied to Windows 3.1 that we have not (yet) succeeded in executing them in the Windows XP environment.

The work described in this paper is part of a larger project which aims to create open-source tools to support assisted emulation and which will greatly expand the set of test cases from those we have discussed. We plan to make all of the data, scripts, and helper code available at the end of the project.

Acknowledgment

This material is based upon work supported by the National Science Foundation under grant No. IIS-1016967. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

6. REFERENCES

- [1] G. Brown. Developing virtual cd-rom collections: The voyager company publications. In *iPRES2011 8th International Conference on Preservation of Digital Objects*, 2011.
- [2] S. Gilheany. Preserving digital information forever and a call for emulators. In *Digital Libraries Asia 98: The Digital Era: Implications, Challenges, and Issues*, 1998.
- [3] M. Guttenbrunner, C. Becker, and A. Rauber. Keeping the game alive: Evaluating strategies for the

- preservation of console video games. *The International Journal of Digital Curation*, 5(1):64–90, 2010.
- [4] A. R. Heminger and S. Robertson. The digital rosetta stone: a model for maintaining long-term access to static digital documents. *Communications of AIS*, 3(1es):2, 2000.
 - [5] Keeping Emulation Environments Portable.
<http://www.keep-project.eu/expub2/index.php>. \AccessedSeptember2011.
 - [6] A. T. McCray and M. E. Gallagher. Principles for digital library development. *Communications of the ACM*, 44(5):48–54, 2001.
 - [7] P. Mellor. CaMiLEON: emulation and BBC doomsday. *RLG DigiNews*, 7(2), 2003.
 - [8] OpenAFS. OpenAFS, 2010. <http://www.openafs.org>. Accessed November 2010.
 - [9] K. Rechert, D. von Suchodoletz, and R. Welte. Emulation based services in digital preservation. In *Proceedings of the 10th annual joint conference on Digital libraries, JCDL*, pages 365–368, 2010.
 - [10] J. Rothenberg. An experiment in using emulation to preserve digital publications. Technical report, Koninklijke Bibliotheek, July 2000.
 - [11] J. Rothenberg. Using emulation to preserve digital documents. Technical report, Koninklijke Bibliotheek, July 2000.
 - [12] D. von Suchodoletz, K. Rechert, J. Schroder, and J. van der Hoeven. Seven steps for reliable emulation strategies -solved problems and open issues. In *7th International Conference on Preservation of Digital Objects (iPRES2010)*, 2010.
 - [13] K. Woods and G. Brown. Assisted emulation for legacy executables. *The International Journal of Digital Curation*, 5(1):160–171, 2010.

Automated Digital Processing at the Bentley Historical Library

Michael Shallcross

Bentley Historical Library
1150 Beal Avenue
Ann Arbor, MI 48108-2113 U.S.A.
shallcro@umich.edu deromedi@umich.edu

Nancy Deromedi

Abstract

Archival processing in the digital era requires traditional steps such as appraisal, arrangement, and description in addition to procedures that ensure the authenticity, integrity, and security of content. Given the labor-intensive nature of manual procedures, the Bentley Historical Library's Digital Curation Division wrote a series of scripts that call various applications and command line utilities and thereby automate key steps in the ingest and processing of born-digital archival materials.

1. Institutional Context

Established in 1935 by the University of Michigan Regents, the Bentley Historical Library serves as the official archives of the university and documents the history of the state of Michigan and the activities of its people, organizations and voluntary associations. The library has successfully managed and preserved digital content since the 1997 accession of former University President James J. Duderstadt's digital desktop. Given the steep increase in born digital and digitized content accessioned by the library in recent years, archivists have sought more efficient and standardized processing procedures. The Andrew W. Mellon Foundation-funded MeMail Project (2010-2011) provided the library with resources to establish a workflow and corresponding policies for the ingest and processing of archival email, but a similar solution was needed for mixed digital content (i.e. Office documents, PDFs, audio and video files, images, etc.). Archivists in the library's Digital Curation Division have advanced the work of the MeMail Project in developing the AutomatedProcessor (or AutoPro), a series of inter-dependent scripts that automates key steps in preparing digital content for long-term preservation and access.

2. Digital Processing as a Concept and

Approach at the Bentley Library

Archival processing in the digital era requires traditional steps such as appraisal, arrangement, and description in addition to procedures that ensure the authenticity, integrity, and security of content. "Digital processing" therefore corresponds to the "generate AIP" function of the Open Archival Information System (OAIS) Reference Model's Ingest entity. After a Submission Information Package (SIP) has been assigned an accession record, digital processing permits archivists to assume intellectual control, establish the integrity of materials, and perform preservation events (i.e. scans for viruses and personally identifiable information, conversion to preservation formats, recording of descriptive and technical metadata, etc.) that transform the SIP into an Archival Information Package (AIP). Bentley archivists initially developed a manual workflow

with more than 40 discrete steps that required the operation of numerous stand-alone applications and saving tool output to various log files. In addition to being highly labor intensive and introducing numerous opportunities for operator error, this approach was daunting for staff without technical expertise. Given these challenges, the Digital Curation division developed AutoPro to fulfill two goals: (1) to make digital processing more efficient by automating key workflow steps and (2) to reduce technical barriers and thereby permit archivists to focus their energies on the traditional archival functions of appraisal, arrangement, and description.

3. Automated Processing: an Overview

AutoPro is comprised of 33 Windows CMD.EXE shell scripts that move content through an 11 step workflow and thereby simplify the operation of more than 20 applications and command line utilities. In addition to providing a framework to guide archivists through the workflow, AutoPro tracks the current processing status, generates log files for all operations, and records PREMIS preservation metadata that will be stored alongside the processed content in a preservation environment. The Windows Command Prompt and Explorer windows function as the main interfaces. Archivists must approve the successful completion of each step and may stop at any point in the workflow and resume their work at a later time.

Immediately after content is accessioned and deposited in the Bentley Library's interim repository (a secure Windows file server), AutoPro runs a virus scan (the University of Michigan employs Microsoft Forefront Endpoint Protection on all work stations) and creates a working backup so the SIP can be restored in case of an error or accidental data loss.

AutoPro then searches for archive files (.ZIP, .TAR, .RAR, etc.); if any are found, a script employs 7-Zip [1] to extract the contents to a directory named after the archive file, with the original file paths preserved. After verifying the extraction's success, AutoPro moves the archive file to a separations directory and records the operations in a log file. The newly extracted content is then searched for additional archive files, from which the contents are extracted, if necessary.

At this stage, AutoPro uses ReNamer [2] to replace spaces and non-alphanumeric characters with underscores in folder and file names and log the original and new names in a comma-separated values (.CSV) file. Next, AutoPro searches for files with missing

or user-supplied extensions, identifies correct extensions with the TrID File Identifier utility [3], and verifies the results with DROID [4]. AutoPro preserves the TrID output (which includes a report on likely file types, based upon the target file's binary signature) in a log file. If an extension is successfully identified, the original and new filenames are recorded in a .CSV file.

In transforming the SIP to an AIP, the Bentley Library relies upon file format conversion as a primary preservation strategy. Based upon the Library of Congress's work on the "Sustainability of Digital Formats" [5] and documentation from the Florida Center for Library Automation and other peer institutions, the library has identified a number of at-risk (i.e. proprietary or potentially obsolete) file formats and developed conversion pathways to sustainable formats with various open source and freeware tools. AutoPro searches for these at risk formats (based upon extension) and then employs the following tools (with digital media and target format in parentheses): ImageMagick (raster images to .TIFF) [6], Ghostscript (.PS and .PDF to .PDF/A; an Adobe Acrobat Preflight droplet verifies if the original PDF meets PDF/A specifications) [7], Inkscape (vector images to .SVG) [8], ffmpeg (audio to .WAV; video to MP4 with H.264 encoding) [9], Aid4Mail (email to .MBOX) [10], and Microsoft Office File Converter (Office files to Open Office XML) [11]. These preservation versions are stored alongside the original and denoted by a suffix consisting of '_bhl-' and the CRC32 hash of the original file (i.e. oralHistoryProject_bhl-0fbcc2cc7.wav). AutoPro also creates a log of all file conversions, including the original and new filenames, timestamp, and conversion software.

In order to protect the identities of record creators and limit its exposure to risk, the Bentley Historical Library has established policies in regard to personally identifiable information (PII) such as credit card numbers and U.S. Social Security numbers. AutoPro thus employs Identity Finder DLP Endpoint [12] to scan for PII. Archivists then use the Identity Finder interface to verify search results and—if true positive hits are found—redact the PII (from Open Office XML and plain text files) or assign appropriate access restrictions to the content.

Archivists then proceed to a more in-depth appraisal and arrangement of content. AutoPro loads data visualizations (such as the distribution of file extensions, date range of content, relative size of directories, etc.) produced by TreeSize Professional to better characterize and launches Quick View Plus (a file viewing program) to rapidly review a wide range of file types for description in finding aids [13]. While reviewing content with Quick View Plus or the Windows Explorer, archivists use a batch file in the right-click context menu to remove superfluous files or folders to a separations directory. Every effort is made to retain the original order of materials, but archivists may group unorganized content in directories or package content in .ZIP files to simplify the management and storage (with such actions recorded in log files). Once the arrangement is established, AutoPro calls DROID to extract technical metadata and generate an MD5 checksum for all content (including files in .ZIP archives). Archivists then use AutoPro to identify series and add descriptive and administrative metadata about the materials; the resulting XML file is used to deposit unrestricted content in Deep

Blue, the University of Michigan's DSpace repository. Finally, AutoPro employs BagIt to transfer a copy of all material to a secure dark archives [14]. Once this step is accomplished, AutoPro cleans the processing directory and temporary files and archivists record the completed digital deposit in the Bentley's collections management database.

4. References

NOTE: all URLs successfully accessed 28 August 2012.

- [1] 7-Zip is an open source file archiving application. For more information see <http://www.7-zip.org/>.
- [2] ReNamer is a freely distributed file renaming tool. For more information, see <http://www.den4b.com/?x=products&product=renamer>.
- [3] TrID is a freely distributed utility that identifies file types based upon a library of over 4,800 binary signatures. For more information, see <http://mark0.net/soft-trid-e.html>.
- [4] DROID is a file identification tool developed by the National Archives (U.K.). For more information, see <http://droid.sourceforge.net/>.
- [5] For more information on the Library of Congress's "Sustainability of Digital Formats," see <http://www.digitalpreservation.gov/formats/index.shtml>.
- [6] ImageMagick is an open source raster image editor. For more information, see <http://www.imagemagick.org/script/index.php>.
- [7] Ghostscript is an open source interpreter for the PostScript language and PDF documents that may be used to convert the latter documents to PDF/A. For more information, see <http://www.ghostscript.com/>.
- [8] Inkscape is an open source vector graphics editor. For more information, see <http://inkscape.org/>.
- [9] ffmpeg is freely available software used for audio and video recording and conversion. For more information, see <http://ffmpeg.org/>. AutoPro utilizes a Windows build available from <http://ffmpeg.zeranoe.com/builds/>.
- [10] Aid4Mail is a proprietary email conversion program. For more information, see <http://www.aid4mail.com/>.
- [11] Microsoft File Convertor is part of the freely available Office Migration Planning Manager. For more information, see <http://www.microsoft.com/en-us/download/details.aspx?id=11454>.
- [12] Identity Finder Data Loss Prevention (DLP) Endpoint is proprietary software that can identify potentially sensitive information. For more information, see <http://www.identityfinder.com/us/Business/IdentityFinder/EnterpriseClient>.
- [13] TreeSize Professional is a proprietary hard disk space and file manager and Quick View Plus is a file viewing utility. For more information, see <http://www.jam-software.com/treesize/> and <https://avantstar.com/>, respectively.
- [14] BagIt is part of an open source set of transfer tools developed by the Library of Congress. For more information, see <http://sourceforge.net/projects/loc-xferutils/>.

Aggregating a Knowledge Base of File Formats from Linked Open Data

Roman Graf

AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
roman.graf@ait.ac.at

Sergiu Gordea

AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
sergiu.gordea@ait.ac.at

ABSTRACT

This paper presents an approach for semi-automatic aggregation of knowledge on computer file formats used to support planning for long term preservation. Our goal is to create a solid knowledge base from linked open data repositories which represents the fundament of the DiPRec recommender system. The ontology mapping approach is employed for collecting the information and integrating it in a common domain model. Furthermore, we employ expert rules for inferring explicit knowledge on the nature and preservation friendliness of the file formats.

Categories and Subject Descriptors

H.3.7 [Digital Libraries]: Systems issues; H.3.5 [On-line Information Services]: Data sharing

1. INTRODUCTION

The core of preservation planning is represented by the file formats used for encoding the digital information. Currently, the information about the file formats lacks a unified well-formed representation in LOD repositories and is only partially available in domain specific knowledge bases (i.e. PRONOM). The activities related to the preservation of digital content are associated with high financial efforts; therefore the decisions about preservation planning must be taken by using rich, trusted, as complete as possible domain knowledge.

The linked open data (LOD) initiative defines best practices for publishing structured data in the Web using a well-defined and queryable format [3]. By linking together and inferring knowledge from different publicly available data repositories (i.e. Freebase, DBpedia, PRONOM) we aim at building a better, more complete characterization of available file formats. In this paper we present the File Format Metadata Aggregator (FFMA) service which implements the proposed approach for building a solid knowledge base supporting digital preservation planning and enactment. FFMA represents the core of the Digital Preservation Recommender (DiPRec) introduced in earlier paper by the authors [1]. The main contributions of this paper consist in: a) proposing and evaluating the approach based on ontology mapping for integrating digital preservation related information from the web; b) using AI models for inferring domain specific

knowledge and for analyzing the preservation friendliness of the file formats basing on the expert models and computation of preservation risk scores.

2. KNOWLEDGE BASE AGGREGATION

One of the main concerns in the design of the FFMA service is the mapping of the semantics between LOD repositories and FFMA domain model. There are two alternatives for mapping file format ontologies: a) by employing ontology matching tools or b) by doing it manually [2]. For development of the FFMA service we chose to perform manual mapping (Fig. 1), due to reduce size of the domain model and the complexity and heterogeneity of the Freebase and DBpedia ontologies.

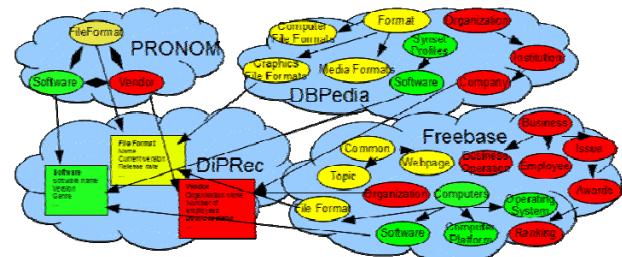


Figure 1. Relationship between data representations.

The underlying domain model consists of three core concepts represented by the File Formats, Software and Vendors. The properties associated to these objects are common for the LOD and PRONOM repositories. The FFMA domain model is aligned with the PRONOM one, which is a reference in the digital preservation domain. Since PRONOM data is not enough documented to cover all computer file formats, and their description is not rich enough for supporting reasoning and recommendations, we collect additional information from LOD repositories and aggregate it in a single homogeneous property based representation in the FFMA knowledge base (Figure 2).

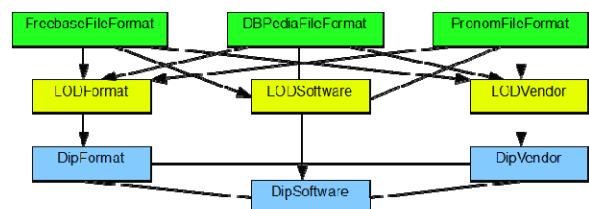


Figure 2. FFMA domain object model overview.

The *FileFormat classes store an index of the individual file format available in each of the external repositories, and they are used for crawling the LOD repositories for relevant information which is stored in LOD* objects. This data is cleaned from

duplications or ambiguities and integrated in the internal representation of file format descriptions which is stored in the DipFormat, DipSoftware and DipVendor classes.

The Domain Knowledge Aggregation is based on the risk analysis model which is in charge of evaluating the information aggregated in the previous step and computing the risk scores over different digital preservation dimensions (e.g. provenance, exploitation context, web compatibility, etc.). A cost based model used for computing the risk scores is designed to provide a simple yet powerful mechanism for definition of expert rules, metrics and classifications used for computing recommendations in DiPRec. A more detailed description and examples on knowledge aggregation process can be found in [1].

3. EVALUATION

The aim of the experimental evaluation is to demonstrate the improvements provided by the proposed approach over the domain specific knowledge base represented by PRONOM. Apart from crawling the information basing on ontology mapping solution we also perform data cleaning in order to remove duplicates and ambiguous information.

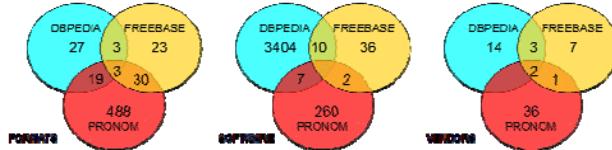


Figure 3. The distribution of objects in LOD repositories.

One of the possible practical user scenarios for FFMA system is the search of software solutions available for creation of the migration plans. The main goal of this scenario is to retrieve rich information on file formats, software and vendors from LOD repositories which allows evaluating the preservation friendliness of software formats.

In experiment we verified our hypothesis that information extraction from additional sources will significantly increase the amount of information available in PRONOM technical registry. The information extraction started with PRONOM (540 formats, 269 software and 39 vendors) was significantly enriched by DBpedia data (52 formats, 3421 software and 19 vendors) and concluded data retrieval with the Freebase (59 formats, 48 software and 13 vendors). In conclusion the FFMA knowledge base stores with ~10% more file formats, about 13 times more software and with 60% more vendors than PRONOM (see Fig. 3). Table 1 demonstrates a significant improvement of the aggregated information broken down to the sample file formats regarding additional knowledge about format versions, software and vendors. E.g. for "GIF" format FFMA comprises the description of 4 of its versions, 6 software tools and 2 vendors more than PRONOM. The multiple data entries in one LOD repository (e.g. two entries for "JPG" format in DBpedia) could be explained either with different versions of the same format or with slightly different names used for the same file format (i.e. identified by same extensions). Given the results presented above, we can demonstrate an important gain when aggregating knowledge from LOD repositories. Moreover, these repositories integrate data from public sources (e.g. like Wikipedia, Stock Market value for Software vendors, Websites of ISO/IETF standards, etc.) which is expected to be grow in time with the support of cross domain information sharing within the given communities.

Table 1. Extracted file format values count in DiPRec classes.

Format	Versions		Software		Vendors	
	PR	FFMA	PR	FFMA	PR	FFMA
TIF	9	19	0	134	0	1
PDF	17	33	14	30	5	6
PNG	3	7	13	28	4	5
GIF	2	6	13	19	4	6
JPG	9	12	13	16	4	5

4. CONCLUSIONS

Within this paper we presented the file format metadata aggregation service which builds a knowledge base with rich descriptions of computer file formats. The service uses semiautomatic information extraction from the LOD repositories, analyzes it and aggregates knowledge that facilitates decision making for preservation planning.

An important contribution of this paper is the usage of the ontology mapping approach for collecting data from LOD repositories. The evaluation of preservation friendliness is based on risk scores computed with the help of expert models. This allows automatic retrieval of rich, up to date knowledge on file formats, reducing so the setup and maintenance costs for the digital preservation expert systems (e.g. DiPRec).

As future work we plan to use additional knowledge sources (e.g. vendor's web sites, further knowledge bases) for extending the knowledge related to the software tools, vendors and their relationship to the existing file formats (which are often missing/incomplete in each of the named repositories). In the same time, we might consider to enhance the modules used for knowledge extraction for inferring further explicit knowledge (e.g. clustering by groups of file formats like text, graphical formats, video, audio file formats, etc.).

5. ACKNOWLEDGMENTS

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

6. REFERENCES

- [1] Gordea, S., Lindley, A., and Graf, R. 2011. Computing recommendations for long term data accessibility basing on open knowledge and linked data. *ACM Trans. Program. Lang. Syst.* 15, 5 (Nov. 2011), 795-825.
- [2] Jain, P., Yeh, P., Verma, K., Vasquez, R., Damova, M., Hitzler, P., and Sheth, A. 2011. Contextual ontology alignment of lod with an upper ontology: A case study with proton. In G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, and J. Pan, editors, *The Semantic Web: Research and Applications*, vol. 6643 of LNCS. Springer Berlin, Heidelberg, 80-92.
- [3] Kurt, B., Colin, E., Praveen, P., Tim, S., and Jamie, T. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD '08 Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, New York, NY, USA, 1247-1249.

Bibliobloggers' Preservation Perceptions, Preferences, and Practices

Carolyn Hank

McGill University

3661 rue Peel, Rm. 210

Montreal, Quebec, H3A 1X1

001-514-398-4684

carolyn.hank@mcgill.ca

Cassidy R. Sugimoto

Indiana University Bloomington

1320 E. 10th Street, LI013

Bloomington, Indiana 47401

001-812-856-2323

sugimoto@indiana.edu

ABSTRACT

The biblioblogosphere comprises the personal, typically professional-oriented publication of blogs by information and library science practitioners, researchers, and educators. This poster presents preliminary findings from a descriptive study examining bibliobloggers' attitudes and preferences for digital preservation, and their respective blog publishing behaviors and blog characteristics influencing preservation action. Findings will be compared to those from an earlier study of blogging scholars from the fields of history, economics, law, biology, chemistry and physics. When considering their dual role as publishers contributing to the scholarly record and, in reflection of their professional roles, work relating to stewardship of this record, bibliobloggers present an exceptional case to extend and better understand the phenomenon of blogging in academe and implications for long-term stewardship of this form.

Categories and Subject Descriptors

I.7.4 [Document and Text Processing]: Electronic Publishing.

H.3.7 [Information Storage and Retrieval]: Digital Libraries.

General Terms

Documentation, Human Factors, Legal Aspects

Keywords

Weblogs, bloggers, digital preservation, scholarly communication

1. PROBLEM AREA

Several neologisms have emerged to reflect academics' blog publications, including bloggeriship in the legal scholarship realm [1] and blogademia for academe in general [2]. The field of information and library science (ILS) has its own: the biblioblogosphere. This neologism, first introduced by Schneider in 2004, as cited by Stephens [3], comprises the institutional publication of blogs of libraries and the personal, typically professionally-oriented publication of blogs by practitioners and ILS-aligned researchers and educators.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

iPRES 2012, October 1–5, 2012, Toronto, Ontario, Canada.

Copyright 2010 ACM 1-58113-000-0/00/0010 ...\$15.00.

While it was recently found that there has been a decline in the number of active blogs within the biblioblogosphere, publication via posting was found to remain stable [5]. But, will these biblioblogs remain stable and available into the future? As noted by Borgman [6], "digital communications leave a trace," but when considering the nature of the blog form as well as the technical, regulatory and social frameworks in which blogging takes place, for how long?

Blogs, as a co-produced medium, represent a mix of code, content and co-producers, comprised not only of bloggers, both the known and the anonymous, but also their readers, service providers, and other contributors. In consideration of the goals of digital preservation – summarized by Caplan [4] as acquiring, describing, interpreting, securing, authenticating, accessing and performing – this multiplicity of co-producers and variety in form and content complicates effective and ethical blog preservation actions. Without deliberate personal or programmatic approaches to the long-term stewardship of these digital communications, the biblioblogs of today may be unavailable into the future.

There has been much work in recent years to meet this need. Research and development in web archiving, such as the Memento project (<http://www.mementoweb.org/>), and blog archiving in particular, such as the BlogForever project (<http://blogforever.eu/>), is an active, ongoing activity. Further, there are many examples of operational blog archiving programs (e.g., The Internet Archive's Way back Machine; the Library of Congress' Legal Blawgs Web Archive), as well as free and fee-based blog back-up services, such as Backupify, BlogBackupr, and BackupMyBlog.

A full treatment of these activities and services is beyond the scope of this short paper. They are referenced here to demonstrate current work in blog preservation, as well as the need to continue to investigate the behaviours and preferences of content creators to further inform and advance programmatic and personal blog preservation activities. This need for blog preservation services was evidenced in an earlier study by one of these authors of 153 scholars' blogging in the areas of history, law, economics, biology, chemistry and physics [7]. Most (80%) agree their blogs should be preserved for public access and use into the indefinite future. It also points to a critical need for guidance, as a majority (61%) report taking some personal action to save all or parts of their respective blogs, though the extent and effectiveness of their approaches varies.

While work is ongoing, there have been no studies to date that specifically examine, in tandem, the blog publishing and preservation behaviors and preferences of bibliobloggers.

Bibliobloggers both contribute to the scholarly record and facilitate stewardship of the scholarly record, whether from an active, hands-on role as library professionals, as educators, preparing the next generation of library professionals, or as researchers, examining compelling issues related to the information use environment. A specific look at bibliobloggers will advance understanding to inform blog preservation activities, and allow a comparison between this particular class of bloggers and those scholars represented in the earlier study.

2. METHODS

Through a mixed-methods approach utilizing qualitative and quantitative data collection activities, this study in progress examines the practices, perceptions, and preferences of select members of the biblioblogosphere – blogging librarians and LIS researchers and educators – and their respective blogs. While the overall study is larger, this poster will report select findings in regard to two specific research questions: 1) how do bibliobloggers perceive their blog in relation to long-term stewardship and, subsequently, who, if anyone, do they perceive as responsible as well as capable for blog preservation; and 2) how do blog characteristics and blogger publishing behaviors and preferences impact preservation action? Findings are drawn from two units of analysis – bibliobloggers and their respective blogs – and three data collection procedures – questionnaires, semi-structured interviews, and blog analysis. Data collection is currently ongoing, anticipated to be completed in early to mid-September 2012.

2.1 Sampling

Bibliobloggers and biblioblogs are identified through purposive sampling. The sampling frames are compiled from two biblioblogs directory listings, LIS Wiki Weblogs and LISZEN, and through a review of faculty listings and websites at all American Library Association accredited graduate programs.

2.2 Data Collection

Bibliobloggers listed to the blogger sampling frame are invited to participate in the questionnaire stage of the study. The questionnaires are comprised of primarily closed-ended questions and organized into ten sections: 1) background; 2) scholarly publishing history; 3) blogging and scholarly communication; 4) blogging activity, identity, and audience; 5) blog publishing behaviors and preferences; 6) blog revision history; 7) blog preservation behavior; 8) blog preservation preferences; 9) other blogging activities; and 10) demographics.

Interview participants will be identified from respondents' returning completed questionnaires. The semi-structured interview schedule, anticipated to take between 20 to 30 minutes to complete, is designed to clarify and further understanding of findings from the questionnaire phase of the study.

For the blog analysis stage, a random sample of blogs listed to the blog sampling frame will be drawn at a 50% sampling ratio. The coding system captures data points across seven categories: 1) authorship; 2) blog elements and features; 3) rights and disclaimers; 4) authority and audience; 5) blog publishing activity; 6) post features; and 7) archiving.

2.3 Analysis

The overall goal of this study is to describe attributes, perceptions, preferences and practices of this particular group of bibliobloggers. Additionally, select findings will be compared to those from the earlier study of other blogging scholars [7]. Quantitative analysis from the questionnaire and blog analysis stages of the study will rely heavily on descriptive measures. Qualitative data from the interview portion of the study will be analyzed using manifest and latent coding techniques, with the goal to identify themes emerging from responses through consideration of frequency, direction and intensity.

3. INTENDED IMPACT

Multiple audiences should benefit from the results of this study, including: 1) bibliobloggers interested in personal preservation of their blog content; 2) organizations with current, piloted or planned digital preservation initiatives who are considering the medium; 3) organizations without planned digital preservation initiatives, in order to inform future, strategic collection policy decisions; and 4) developers and researchers working on the area of web and blog archiving. Findings from this research are intended to inform decision-making, today, on selection and appraisal of biblioblogs for access and use into the future.

4. ACKNOWLEDGMENTS

This work is supported through the OCLC/ALISE Library and Information Science Research Grant Program (LISRGP) for the study, "The Biblioblogosphere: A Comparison of Communication and Preservation Perceptions and Practices between Blogging LIS Scholar-Practitioners and LIS Scholar-Researchers."

5. REFERENCES

- [1] Caron, P.L. 2006. Are scholars better bloggers? Bloggership: how blogs are transforming legal scholarship. *Washington University Law Review*. 84, 5 (Nov. 2007), 1025-1042.
- [2] Saper, C. 2006. Blogademia. *Reconstruction*. 6, 4 (Nov. 2007), 1-15.
- [3] Stephens, M. 2008. The pragmatic biblioblogger: Examining the motivations and observations of early adopter librarian bloggers. *Internet Reference Services Quarterly*. 13, 4 (Dec. 2008), 311-345.
- [4] Caplan, P. 2008. The preservation of digital materials. *Library Technology Reports*. 44, 2 (Feb.-Mar. 2008), 5-38.
- [5] Torres-Salinas, D., Cabezas-Claijo, A., Ruiz-Perez, R., and Lopez-Cozar, E. 2011. State of the library information science blogosphere after social networks boom: A metric approach. *Library Information Science Research*. 33, 2 (Feb. 2011), 168-174.
- [6] Borgman, C.L. 2008. *Scholarship in the digital age: Information, infrastructure and the Internet*. MIT Press, Cambridge, MA.
- [7] Hank, C. 2011. *Scholars and their blogs: Characteristics, preferences and perceptions impacting digital preservation*. Doctoral Thesis. UMI Order No. 4356270, University of North Carolina at Chapel Hill.

Poster ‘Preservation through access: the AHDS Performing Arts collections in ECLAP and Europeana’

Perla Innocenti

CCA, University of Glasgow
8 University Gardens
Glasgow, UK, G12 8QH
perla.innocenti@glasgow.ac.uk

John Richards

CCA, University of Glasgow
8 University Gardens
Glasgow, UK, G12 8QH
john.richards@glasgow.ac.uk

1. INTRODUCTION

This poster provides an overview of the ongoing rescue of valuable digital collections that had been taken down and consequently lost to general access.

The University of Glasgow was home to the Arts and Humanities Data Service Performing Arts (AHDS Performing Arts) [1], one of the five arts and humanities data centres that constitute the Arts and Humanities Data Service (AHDS). Since 1996 AHDS supported the creation, curation, preservation and reuse of digital materials for the UK Arts and Humanities research and teaching community. AHDS Performing Arts, based in Glasgow, supported research, learning and teaching in music, dance, theatre, radio, film, television, and performance for thirteen years. Working with the AHDS Executive, relevant performing arts collections have been ingested, documented, preserved, and where possible made available via the AHDS Cross Search Catalogue and Website to researchers, practitioners, and the general public. Furthermore strong relationships were developed with research and teaching community upon a scoping study investigating user needs [2].

In 2007 the co-funders of the AHDS - Arts and Humanities Research Council (AHRC) for the UK and the Joint Information Systems Committee (JISC) - withdrew their funding. A detailed risk assessment report was produced in response to the withdrawal of core funding [3], but to no avail. When the AHDS funding stopped, online access to these cultural resources eventually became discontinued [4].

In 2010, the School of Culture and Creative Arts at the University of Glasgow joined the EU-funded ECLAP project to ensure that at least part of these resources could be accessible for the long term by scholars and practitioners in the performing arts arena, and by the general public. Below we briefly describe the ECLAP project, the AHDS Performing Arts collections progressively available through it and some thoughts on providing preservation through access for this type of digital cultural resources.

2. ECLAP project

ECLAP (European Collected Library of Artistic Performance, www.eclap.eu/) is an EU-funded Best Practice Network infrastructure and a service portal providing a large, collaborative and multilingual online library for performing arts institutions and users in Europe and beyond. Through its portal and services ECLAP aims at enriching and promoting performing arts heritage and culture, and supporting advances in learning and research in the field of performing arts. The project Consortium, led by Prof. Paolo Nesi at DSI University of Florence, is composed by European leading national performing arts institutions, universities and research institutes [5].

ECLAP offers a wide range of innovative solutions and tools to support performing arts institutions in managing, providing access to and disseminating their online collections to a large number of users. This initiative is bringing together hundreds of thousands of Europe’s most relevant performing arts content (previously often inaccessible online), including collections on theatre, dance, music, cinema and film, performances, lessons, master classes, teaching, festival, costumes, sketches, production materials, lyrics, posters, locations. File formats include video and audio files, documents, images, animations, playlists, annotations, 3D, interactive content, e-book, slides. ECLAP is fully integrated with Europeana, the portal endorsed by the European Commission providing a single access point to millions European cultural and scientific heritage digital objects [6].

3. AHDS PERFORMING ARTS COLLECTIONS IN ECLAP

3.1 Value and challenges for performance digital content

Performing arts represent a valuable cultural resource, and also an important economic sector in the Creative Industries. According to the Department of Culture, Media and Sport’s annual analysis of the impact of the creative industries on British economy [7], the Music & Visual and Performing Arts sector is the Creative Industries largest contributor, employing in 2009 some 300,000 people and contributing £3.7 billion of Gross Value Added (GVA) to the economy. As Seamus Ross noted, for the performing arts community ‘the process of creative activity is itself a core deliverable. While we need to document performance, giving preference to print and text as a way to describe performance is not adequate’ [2, Preface]. However performing arts outputs are often constituted by ‘rich audiovisual resources that are complex to preserve from both a technical and intellectual point of view’ [3]. Furthermore, typically performance collections have an online bibliographic catalogue, through which users can sometimes access a limited amount of digital content, often restricted to textual materials.

3.2 Migration challenges

Considering the complexity, heterogeneity and diversity of types and formats of AHDS PA collections, their migration from one environment (especially a lost one) included a number of organisational, semantic and technical challenges. Some examples are mentioned here.

Conflicts in selection and appraisal. We had originally selected for ingestion into ECLAP and Europeana about half of the sixty-one AHDS PA deposited collections, that is the collections made

available for public dissemination. Our selection paid particularly attention to the collections that were categorized as ‘Medium risk’ and ‘Relatively high risk’ in the risk assessment conducted in 2007, when AHDS withdrew its funding [3]. However Europeana does not accept intellectually produced content derived from research in the form of databases, as they are considered at the same level of library catalogues despite the substantial intellectual endeavor behind them. After a failed attempt to negotiate the ingestion of such resources, we had to further reduce the selection of AHDS PA collections to eleven (Adolphe Appia at Hellerau, Anatomical Exoskeleton, Cheap Flight Show, Citywide, Designing Shakespeare, Fabulous Paris, Five Centuries of Scottish Music, From Jayaprana to the Abduction of Sita, Imago, Imago lecture notes, King Lear Performance Photographs).

Physical rescue of the collections. The AHDS PA collections, which were originally physically stored at the University of Glasgow, had been moved to King’s College in London during the lifetime of the project. The rescue of the collections included an onsite journey to King’s College, where the files were physically stored but no longer made available online, and three attempts over a period of time longer than expected to migrate the large amount of data onto physical storage media and in a secure online environment. Validation procedures were performed to ensure data integrity not only for discrete entities via checksums, but also manually to safeguard the integrated nature of these collections and the context provided to individual objects.

Metadata transfer. The AHDS PA metadata format precedes the Europeana data model by about a decade. AHDS PA metadata format was an in-house metadata structure that has come to be known as the Common Metadata Framework (CMF), developed by Malcolm Polfreman and adopted for all AHDS collections in order to display metadata from all its centres via a single interface. CMF can be mapped to a looser metadata structures such as the Dublin Core underlying the Europeana Data Model (EDM). But while initially EDM was very library-oriented (Dublin Core plus a few fields), the current EDM version draws more on the CIDOC CRM model, predominant in museum. Furthermore, Europeana attempts to enrich data by adding standard multilingual terms and references from thesauri or controlled vocabularies. However there is currently no standard classification scheme for performing arts; the effort of ECLAP Consortium to produce and then use a new classification proved rather challenging.

3.3 Benefits for AHDS Performing Arts collections

Rather than a scattered group of host institution’s websites, the ECLAP portal represents a central point which allow European-wide institutional collections be described, cross-searched and compared. Collection holders can maintain rights and access restrictions as appropriate, while benefiting from the greater visibility provided by the search, discover, retrieve and access via ECLAP and Europeana. In addition, this portal offers a number of search, content enriching, social networking and distribution services through which we can ‘market’ the AHDS PA collections. Services include search and retrieve of multimedia content via a multilingual interface in 21 languages; possibility to enrich, contextualize, annotate, rate and aggregate content;

possibility to share content, including collaborative indexing and creation of forums; distribution and access all content also via mobile devices; e-learning support .

In terms of sustainability, the portal centralised access point, ECLAP strategic planning and its Europeana connection also provide individual performing arts collections with a supporting environment for the long term curation and preservation of their digital assets. For example recent developments in Europeana, the cultural and scientific heritage portal endorsed by the European Commission, are promising in terms of digital preservation (the ASSETS project [8] includes Europeana software services for preparing ground for digital preservation).

4. CONCLUSIONS AND OUTLOOK

In this poster we have provided an overview of the University of Glasgow efforts, via the ongoing EU-funded ECLAP project, to ensure that at least part of the AHDS Performing Arts collections can continue to be accessible for the long term by scholars and practitioners in the performing arts arena, and by the general public. It is our hope that in doing this we will contribute to facilitating discovery, access, understanding, and use of digital performing art resources for current and future generations.

5. ACKNOWLEDGMENTS

ECLAP is co-funded by the European Union ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, Theme CIP-ICT-PSP.2009.2.2, Grant Agreement N° 250481. We are grateful to the AHDS Performing Arts Principal Investigator, Prof. Seamus Ross, for having supported the use of selected AHDS collections in the ECLAP project.

6. REFERENCES

- [1] AHDS Performing Arts, <http://www.ahds.ac.uk/performingarts/index.html>.
- [2] Abbott, D. and E. Beer, Getting to Know Our Audience: AHDS Performing Arts Scoping Study, 2005, www.ahds.ac.uk/performingarts/pubs/scoping-study-2006.pdf
- [3] Jones, S., Abbott, D., and Ross, S., Risk Assessment for AHDS Performing Arts Collections: A Response to the Withdrawal of Core Funding, AHDS Performing Arts, 2007, www.hatii.arts.gla.ac.uk/ahdspa_collection_risk_assessment.pdf
- [4] AHDS Performing Arts Collections listed at <http://www.ahds.ac.uk/performingarts/collections/index.htm> are no longer available within this website.
- [5] ECLAP Partners, <http://www.eclap.eu/partners/>.
- [6] Europeana, www.europeana.eu/portal/.
- [7] Department for Culture, Sports and Media, Creative Industries Economic Estimates Full Statistical Release, 8 2011, www.culture.gov.uk/images/research/Creative-Industries-Economic-Estimates-Report-2011-update.pdf.
- [8] ASSETS (Advanced Service Search and Enhancing Technological Solutions for the European Digital Library), <http://62.101.90.79/web/guest/welcome>

A Digital Repository Year: One Museum's Quest for the Basics

Paula Jabloner
Computer History Museum
1401 N. Shoreline Blvd.
Mountain View, CA 95128 USA
1 (650) 810-1016
pjabloner@computerhistory.org

Katherine Kott
Katherine Kott Consulting
katherinekott@katherinekott.com

Keywords

Digital Repository, Digital Preservation

1. INTRODUCTION

The Computer History Museum (CHM) had its own mini deluge of digital data. Our in-house produced high definition oral histories, lectures and exhibition videos were usurping our available server space at over 60 terabytes, with another 10 terabytes of historic digital artifacts including images and software. With the aid of grant funds from Google.org, CHM took on the work of creating a prototype digital repository in one year. The digital repository working group is excited about the possibilities the new repository represents for expanding our digital collection while putting the Museum in the forefront of small cultural institutions creating digital repositories and we hope to share what we have learned with other similar organizations.

We needed to find solutions that could be managed by an organization of our size (less than 50 employees), yet offered the flexibility to handle the wide range of content we collect. The assumptions we used were based on the museum's immediate needs and time constraints.

They include:

- The digital repository will use existing tools and systems
- CHM will not add staff to build a custom solution
- Open source software will play a significant part in the digital repository management solution
- The preservation layer will be built on top of common commodity storage components that are modular and extensible
- The creation of a digital repository is an on-going commitment by CHM

So far we have created policies, have selected and are implementing software and storage hardware. We are now in the fourth quarter of our year odyssey and are ingesting a small sample set of digital objects to test the prototype. We achieved this in a year carefully defined by quarterly phases.

2. CONFRONTING THE PROBLEM: PREPARATION

Here we:

- Defined the problem and the catalyst
 - 60 terabytes of Museum produced high definition (HD) video
 - with no sustainable back-up or preservation methods
- Cultivated permanent stakeholders from senior management and the Board of Trustees
- Engaged cross-departmental working group of four digital preservationists

3. CREATING THE PROBABLE SOLUTION: PLANNING

Here we:

- Hired a digital repository consultant
 - The 'authority' she gave the project in the eyes of the stakeholders was invaluable
- Created a concise Project Charter defining scope, objectives, roles, roadmap, and assumptions
- Surveyed the Museum's 'archival' digital objects
- Performed a current literature survey and wrote best practices guide

4. CURATION: POLICY & FRAMEWORK

Here we:

- Wrote digital repository management software functional requirements
- Surveyed and test drove open source digital repository management software
 - Selected *Archivematica*
- Recruited and hired a storage infrastructure consultant
- Explored storage options, configurations, and pricing
- Completed a policy document

5. COMPLETING THE PROTOTYPE (ONGOING)

We are:

- Testing the DIY storage infrastructure (hardware & software stack)

- installing storage infrastructure and *Archivematica*
- Ingesting test digital objects while creating procedures document
- Writing a 5-year sustainability plan
- Exploring avenues for year two funding for ingest, full deployment, and prototyping an on-line interface

6. STORAGE INFRASTRUCTURE

We firmly believe the straightforwardness of the storage infrastructure will guarantee the sustainability of the digital objects entrusted in its care. This DIY infrastructure is comprised of:

- Working space for backups of non-ingested digital objects and archive space on the same infrastructure totaling 256 terabytes of raw storage.
- Supermicro storage using 3 TB SATA drives running either *NexentaStor* or *FreeNAS*.
- Two backup servers and Supermicro storage with fewer terabytes. One on-site and the other at our off-site storage facility.
- LTO 5 tape backups with tapes stored in both locations.
- Main server running Archivematica, rsync and other software.

7. CONCLUSION

The working group is excited about the possibilities the new digital repository represents for expanding our digital collection while putting the Computer History Museum in the forefront of small cultural institutions creating digital repositories.

Our lessons learned are that the three most important ingredients were setting the correct expectations from the beginning, adequate planning with an emphasis on quarterly results, and having the right team in place that was both dedicated to the project and with the right mix of experience, talents and abilities. This truly took a team effort.

8. ACKNOWLEDGEMENTS

The Computer History Museum's digital repository work was supported by a generous grant from Google.org. In addition to the authors the Museum's digital repository working group includes:

- Al Kossow, Software Curator
- Ton Luong, IT Manager
- Heather Yager, Digital Media Archivist

PDF to PDF/A: Evaluation of Converter Software for Implementation in Digital Repository Workflow

Jamin Koo

University of Michigan
4322 North Quad
105 S State St
Ann Arbor, MI 48109
jaminkoo@umich.edu

Carol C.H. Chou

Florida Digital Archive,
Florida Virtual Campus - Gainesville
5830 NW 39th Ave.
Gainesville, FL 32606
002-1-352-392-9020
cchou@ufl.edu

ABSTRACT

PDF/A is a version of Portable Document Format backed by ISO standard that is designed for archiving and preservation of electronic documents. Many electronic documents exist in PDF format. Due to its popularity, the ability to convert an existing PDF into a conforming PDF/A file is as important, if not more, as being able to produce documents in PDF/A format in digital preservation. In recognition of this fact and encouraged by growing interest from its affiliates, the Florida Digital Archive (FDA) conducted an evaluation of several of the PDF to PDF/A converter applications, the result of which is reported in this paper. There is room for interpretation in the ISO standards concerning PDF/A, which can be manifest in the development of software. In selecting a PDF to PDF/A converter product, reliability of the outcome in terms of PDF/A compliance must be established along with functionality. The goal of this paper is not to rank or promote the software evaluated, but rather to document the FDA's evaluation process and present the results in such a way that they provide insight into challenges and potential drawbacks during similar evaluation or implementation.

1. INTRODUCTION

The FDA has been in production since 2005. As of 2012, the FDA has over a hundred thousand PDFs in its archive with the presence of all PDF versions from 1.1 to 1.7 where 90 percent of them are version 1.4. Though FDA has encouraged its affiliates to submit PDF/A, less than 1 percent of its PDF archive is identified to be PDF/A-1b using JHOVE's PDF/A-1 validation¹.

To ensure the long-term preservation of its PDFs in the archive, FDA conducted a study to select a PDF to PDF/A conversion application as part of its PDF format normalization strategy in the summer of 2012. The ultimate goals will be 1) to provide better PDF/A validation than the existing one provided by JHOVE; and 2) to normalize all non-PDF/A PDFs in the archive into at least PDF/A-1b.

Eight products currently available in the market were identified from the PDF/A Competence Center on the PDF Association website, of which three were selected for in-depth evaluation after a thorough review of product specifications. Most selection criteria have general applicability, such as the ability to fix unembedded fonts and device-dependent color spaces; however, some requirements, such as Linux support and command line operation, were FDA specific. This paper evaluates PDF/A validation and conversion features of the three products selected,

which are pdfaPilot CLI v3.1.159, 3-Heights PDF to PDF/A Converter v4.0.9.0 and PDF/A manager v5.80. The desktop version of pdfaPilot was also used but for troubleshooting purposes only.

2. VALIDATION

The Bavaria Report [1] is a thorough analysis of PDF/A validation products published in 2009, which included two of the three products assessed in this study. Given the age of the report, the FDA decided to do a preliminary validation testing on the most recent version of all three products using the same test files on which the Bavaria Report was based. The Isartor testsuite² was excluded as the two products already showed 100% compliance in the Bavaria Report on Isartor testsuite.

Table 1: Validation Testing

	Total	False alarm	Miss	Accuracy
pdfaPilot	80	0	8	90%
3-Heights	80	17	4	74% or 95%
PDF/A Manager	80	0	7	91.3%

Note that 3-Heights flagged 17 conforming PDFs as invalid due to embedded fonts declared in the form fields when no form field was visible in the document. PDF Tools, the maker of 3-Heights, confirmed this as a bug that would be addressed in future releases. With this corrections, the accuracy of 3-Heights goes from 74% to 95%.

The differences in accuracy were not enough to indicate superior performance by any of the products on PDF/A validation. However, pdfaPilot produced notably better and more detailed error reporting out of the three.

3. CONVERSION, CROSS-VALIDATION

The conversion testing for each product was based on 203 PDFs chronologically sampled from the FDA archive, which all three products identified as not PDF/A compliant during initial validation. The conversion testing includes pre-conversion validation, conversion, self-revalidation on output files, and cross-revalidation by the other two products. All conversion operations were performed per the PDF/A-1b compliance level.

The Initial Conversion Success Rate and Actual Conversion Success Rate in Table 2 represent the percentage of successful conversions based on post-conversion self-validation and the success rate after an in-depth review of conversion logs and error reports, respectively. False positives (non-compliant output files

¹ JHOVE does not parse the contents on streams, so it cannot determine PDF/A conformance to the degree required by ISO 19005-1.

² Isartor testsuite is a set of files by PDF/A competence center to check software conformance on PDF/A-1 standard.

that passed self-validation) were identified through verification of errors and, in some cases, visual inspection of the files.

Table 2 Conversion Success Rate by Product

	Initial Conversion Success Rate	Actual Conversion Success Rate
pdfaPilot	79.7%	79.7% (--)
3-Heights	89.6%	84.2% (↓)
PDF/A Manager	92.1%	83.7% (↓)

The slightly higher conversion success rates shown by 3-Heights and PDF/A Manager can be attributed to the way these products handle encryption and embedded files. While pdfaPilot required the input files be free of these inhibitors, 3-Heights and PDF/A Manager "fixed" the problem by simply removing such items. However, in the case of non-working bookmarks, 3-Heights and PDF/A Manager flagged them with invalid destination errors, whereas pdfaPilot ignored them and completed the conversion without fixing the bookmarks.

Table 3: Conversion Failures by Product

	pdfaPilot	3-Heights	PDF/A Mgr
Environment Issues	14 (33%)	12(38%)	0
Embedded files	6(17%)	0	0
Encrypted	4(10%)	0	0
Problem PDF	17(40%)	9(28%)	16(38%)
False Positive	0	11(34%)	17(52%)

The conversion errors were grouped into four categories: 1) environment issues, such as fonts and color profiles availability; 2) embedded files in input PDF files; 3) encryption; and 4) other problems in input PDF files including but not limited to syntax and metadata issues. The false positive results from 3-Heights and PDF/A Manager were due to the products failing to detect mostly font-related (environment) and syntax/metadata (other) issues. Both products converted a few files with mis-rendered characters due to a Symbol-Italic font that was un-embedded and unavailable in the system for a fix, resulting in visual differences between the original and the output files (e.g. "beta" italic character appearing as a rectangle). Many of the false positives by PDF/A Manager resulted from the product failing to detect and/or fix XMP issues (e.g. missing XMP packet headers) per XMP Specification 2004 [4] referenced by ISO 19005-1 [2].

4. CHALLENGES

The environment issues are directly tied to the rendering and usability of the files. Even a single missing or mis-rendered glyph, as seen in some false positive files by 3-Heights and PDF/A Manager, can be difficult to detect without proper flags and warnings and have a devastating impact especially in PDFs with scientific data. One of the biggest potential roadblocks in dealing with fonts and color profiles is the rights issues. There are ways to circumvent possible copyrights infringement through font substitution but some specialized fonts may prove to be difficult not only to procure but also to use in PDF/A conversion, as their makers can prohibit embedding of fonts.

Handling of inhibitors like embedded files and encryption also needs to be considered in PDF to PDF/A conversion. While

embedded files can become non-issue per later PDF/A standards, encryption of any type can hinder long-term preservation efforts including the conversion to PDF/A. Indiscriminate removal of encryptions or embedded files should be employed with caution because of potential adverse effects that may not be immediately evident, although the ability to remove non-critical encryptions may indeed prove useful to some institutions.

As thorough as the standards and documentations for both the PDF and PDF/A formats are, there is room for interpretation in determining the PDF/A compliance, between different documentations in particular. A pertinent example concerns the opposite positions that PDF Tools (maker of 3-Heights) and callas software (maker of pdfaPilot) take regarding non-working bookmarks. While the invalid destination error is a legitimate error per PDF 1.4 reference [3], there is no specific provision about bookmarks and destinations in ISO 19005-1 [2], which is why callas software does not consider the invalid destination error severe enough to stop or fail conversion for even when pdfaPilot cannot fix or restore the bookmark functionality.

5. CONCLUSION

Establishing reliability and accuracy of PDF/A converter software is not as clear-cut as one might wish, due to the variables involved and challenges demonstrated above. Purely quantitative assessment of the product performance has proven difficult even with adjusted statistics based on extensive analysis of errors. Given the complexity of PDF/A compliance requirements and the automatic fixes applied by the products during the conversion process, which will only grow more sophisticated as technology advances, the two most apparent differentiators are 1) the level of documentation and reporting capabilities of the product; and 2) the access to knowledgeable support staff. For these reasons, this study found pdfaPilot more reliable than the other two products.

6. FUTURE WORK

PDF/A-2 accommodates more features such as embedded files, JPEG 2000, transparency, etc. In addition, to yield higher successful conversion, pdfaPilot also provides a "force-conversion" feature that can convert problem pages into images with invisible text, still allowing marking, searching and copying. The FDA hope to find some resources in the future to continue the PDF to PDF/A conversion testing with a focus on PDF/A-2 and the pdfaPilot's force-conversion feature.

7. ACKNOWLEDGEMENTS

The authors wish to thank Priscilla Caplan and Lydia Motyka for providing insightful feedback and support on this project.

8. REFERENCES

- [1] PDFlib. May 4, 2009. Bavaria Report on PDF/A Validation Accuracy,
<http://www.pdflib.com/fileadmin/pdflib/pdf/pdfa/2009-05-04-Bavaria-report-on-PDFA-validation-accuracy.pdf>
- [2] International Standard Organization, Dec 1, 2005, Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1)
- [3] Adobe Systems Incorporated, Dec 2001, PDF Reference version 1.4
- [4] Adobe Systems Incorporated, Jan 2004, XMP Specification,
<https://www.aiim.org/documents/standards/xmpspecification.pdf>

On the Complexity of Process Preservation: A Case Study on an E-Science Experiment

Rudolf Mayer
Secure Business Austria
Vienna, Austria
rmayer@sba-research.at

Stephan Strodl
Secure Business Austria
Vienna, Austria
sstrodl@sba-research.at

Andreas Rauber
Secure Business Austria
Vienna, Austria
arauber@sba-research.at

ABSTRACT

Digital preservation of (business) processes is an emerging topic in Digital Preservation research. Information technology driven processes are complex digital objects, living in a broad context of aspects relevant to their preservation. In this poster, we detail the broad environment of one sample process from the domain of E-Science, a genre classification experiment in the domain of Music Information Retrieval. We show the magnitude of aspects involved, on technology as well as organisational, legal and other aspects.

General Terms

Process Preservation, Case Study, E-Science

1. INTRODUCTION

Preservation of information technology driven business and scientific processes is an emerging topic in Digital preservation research. These processes are complex digital objects, themselves including and using many other digital objects along the process execution. In this poster, we want to demonstrate on how complex the context of an even rather simple scientific workflow with a limited number of processing steps may become. We show tool support for defining and visualising this context.

2. MUSIC CLASSIFICATION PROCESS

The specific process used in our case study is a scientific experiment in the domain of Music Information Retrieval, where the researcher performs an automatic classification of music into a set of predefined categories. This type of experiment is a standard scenario in music information retrieval research, and is used with many slight variations in set-up for numerous evaluation settings, ranging from ad-hoc experiments to benchmark evaluations such as e.g. the MIREX genre classification or artist identification tasks [1].

The experiment involves several steps; a model of the process in BPMN 2.0, is depicted in Figure 1. First, music data

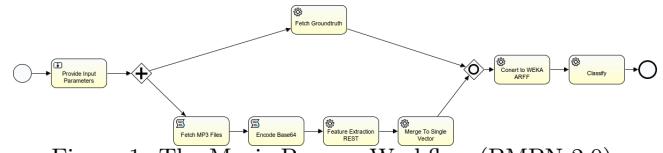


Figure 1: The Music Process Workflow (BPMN 2.0)

is acquired from sources such as benchmark repositories or, in more complex settings, online content providers. In parallel, genre assignments for the pieces of music are obtained from ground truth registries, frequently from websites such as Musicbrainz.org. Tools are employed to extract numerical features describing certain characteristics of the audio files. In the case of the experimental set-up used in this example E-Science process, we assume a more complex scenario where an external web service is used to extract such features. This forms the basis for learning a machine learning model using the WEKA machine learning software, which is finally employed to predict genre labels for unknown music. Further, several scripts are used to convert data formats and other similar tasks. The process described above can be seen as prototypical from a range of E-Science processes, consisting both of external as well as locally available (intermediate) data, external web services as well as locally installed software used in the processing of the workflow, with several dependencies between the various components.

Figure 2 gives an overview on the elements identified as relevant aspects of the business process context, and their relations to each other; we will describe some of these elements below. As the scientific experiment is a process mostly focusing on data processing, a significant amount of the identified aspects are in the technical domain – software components directly used in the processing steps (and their dependencies), external systems such as the web service to extract the numerical audio features from, or data exchanged and their format and specification. However, also *goals* and *motivations* are important aspects, as they might heavily influence the process. As such, the motivation for the providers of the external systems is relevant, as it might determine the future availability of these services. Commercial systems might be more likely to sustain than services operated by a single person for free. Another important aspect in this process are *licenses* – depending on which license terms the components of our process are released under, different options of preservation actions might be available or not. For closed-source, proprietary software, migration to a new execution platform might be prohibited.

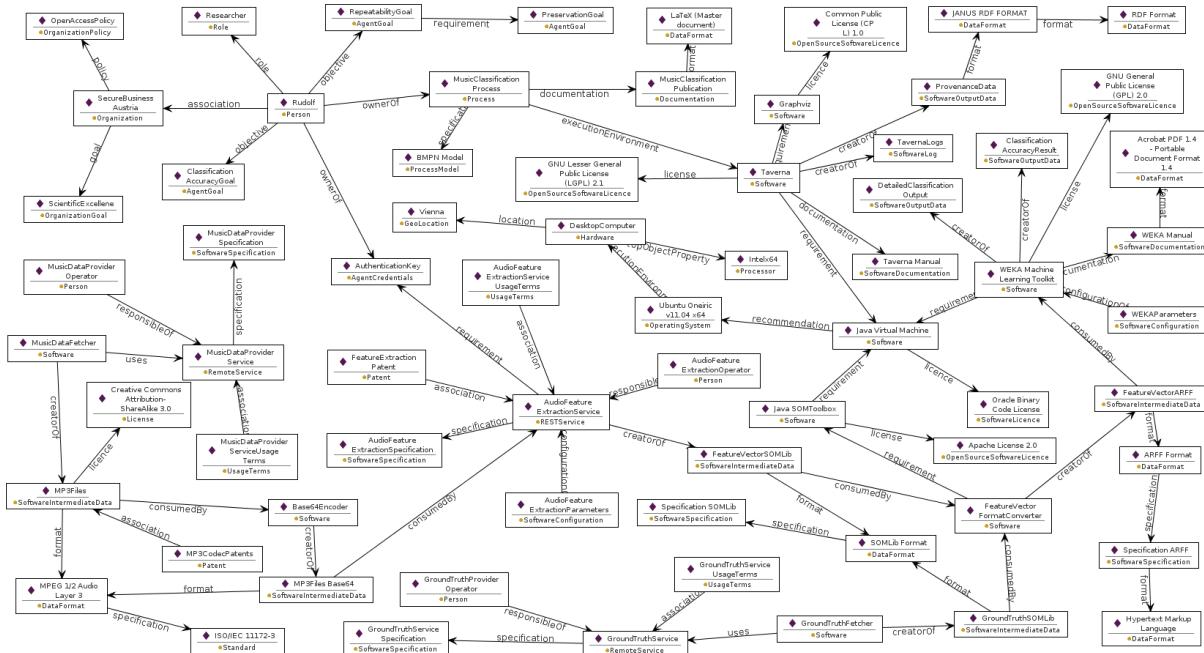


Figure 2: Relevant aspects identified in the scientific workflow

A central aspect in the scientific process is the *AudioFeature-ExtractionService*, i.e. the remote web-service that provides the numeric representation for audio files. The service needs as input files encoded in the *MP3 format* (specified by the *ISO standard 11172-3*). More specifically, as they are binary files, they need to be further encoded with *Base64*, to allow for a data exchange over the HTTP protocol. The web-service accepts a number of parameters that control the exact information captured in the numeric representation; they are specified in the *AudioFeatureExtractionSpecification*, which is authored as a PDF document. The specification further provides information on how the extraction works (i.e. a coarse *documentation* of the signal processing steps applied to obtain the final result). The operator of the web-service provides the service for free, but requires *authorization via a key* that, which is granted to a *person*, and can't be shared under the *usage terms*. The service returns the numeric description as ASCII file, following the *SOMLib format specification*, which is authored in *HTML*.

As a software component used locally, the *WEKA* machine learning toolkit requires a *Java Virtual Machine* (JVM) platform to execute. The JVM in turn is available for many operating systems, but has been specifically tested on a Linux distribution, *Ubuntu, version “Oneiric” 11.04*. WEKA requires as input a feature vector in the *ARFF Format*, and a set of *parameters* controlling the learning algorithm. These parameters are specified in the *WEKA manual*, available in *PDF Format*. As output result, the numeric performance metric “accuracy” is provided, as well as a textual, detailed description of the result. WEKA is distributed under the terms of the open-source GNU Public License (GPL) 2.0, which allows for source code modifications.

After this experimentation process, a subsequent process of result analysis and distillation is normally performed, taking input from the experiment outcomes, and finally leading to

a publication of the research in the form of e.g. a conference or journal *publication*. Here it is modelled as a single information object (the paper written in *LaTeX*) connected to the process, and thus to all data and processing steps that led to the results published. It might also be modelled as a process in its own, specifically if a paper reports on meta-studies across several experiment runs.

Tool support for automatically extracting, manually creating and viewing such process context has been implemented, and will be demonstrated.

3. ACKNOWLEDGMENTS

Part of this work was supported by the FP7 project TIM-BUS, partially funded by the EU under the FP7 contract 269940.

4. REFERENCES

- [1] Music Information Retrieval Evaluation eXchange (MIREX). Website. <http://www.music-ir.org/mirex>.

The Community-driven Evolution of the Archivematica Project

Peter Van Garderen
President, Artefactual Systems, Inc.
202-26 Lorne Mews
New Westminster, BC, Canada
1.604.527.2056
peter@artefactual.com

Courtney C. Mumma
Systems Archivist, Artefactual Systems Inc.
202-26 Lorne Mews
New Westminster, BC, Canada
1.604.527.2056
courtney@artefactual.com

ABSTRACT

In this paper, we discuss innovations by the Archivematica project as a response to the experiences of early implementers and informed by the greater archival, library, digital humanities and digital forensics communities. The Archivematica system is an implementation of the ISO-OAIS functional model and is designed to maintain standards-based, long-term access to collections of digital objects. Early deployments have revealed some limitations of the ISO-OAIS model in the areas of appraisal, arrangement, description, and preservation planning. The Archivematica project has added requirements intended to fill those gaps to its development roadmap for its micro-services architecture and web-based dashboard. Research and development is focused on managing indexed backlogs of transferred digital acquisitions, creating a SIP from a transfer or set of transfers, developing strategies for preserving email, and receiving updates about new normalization paths via a format policy registry (FPR).

General Terms

Documentation, Performance, Design, Reliability, Experimentation, Security, Standardization, Theory, Legal Aspects.

Keywords

archivematica, digital preservation, archives, OAIS, migration, formats, PREMIS, METS, digital forensics, agile development, open-source, appraisal, arrangement, description, acquisition

1. INTRODUCTION

The ISO 14721-OAIS Reference Model [1] gave the archives community a common language for digital archives architectures. One such architecture is the Archivematica suite of tools which

was based on an extensive requirements analysis of the OAIS functional model [2]. The Archivematica project is nearing its first beta release. Project partners and independent implementers have been testing alpha releases using real-world records. These activities have identified some OAIS requirement gaps for digital archives systems.

The project has found that, while it serves as an excellent foundation and framework for long-term preservation strategies, the OAIS model proves inadequate to address some functions unique to archives. In particular for the areas of appraisal, arrangement, description, and preservation planning there were clear gaps between the model and the way that archivists actually process records. The Archivematica project has added requirements to its development roadmap to fill those gaps in its micro-services architecture and web-based dashboard. Other research and development is focused on managing a backlog of indexed digital acquisitions, creating a Submission Information Package (SIP) from a transfer or set of transfers, developing strategies for preserving email, and receiving updates about new normalization paths via a format policy registry (FPR).

2. ABOUT THE ARCHIVEMATICA PROJECT

The Archivematica system uses a micro-services design pattern to provide an integrated suite of free and open-source software tools that allows users to process digital objects from ingest to access in compliance with the ISO-OAIS functional model [3]. It allows archivists and librarians to process digital transfers (accessioned digital objects), arrange them into Submission Information Packages (SIPs), apply media-type preservation plans and create high-quality, repository-independent Archival Information Packages (AIPs). Archivematica is designed to upload Dissemination Information Packages (DIPs) containing descriptive metadata and web-ready access copies to external access systems such as DSpace, CONTENTdm and ICA-AtOM. Users monitor and control the micro-services via a web-based dashboard.

A thorough use case and process analysis identified workflow requirements to comply with the OAIS functional model. Through deployment experiences and user feedback, the project has expanded beyond OAIS requirements to address analysis and arrangement of transfers into SIPs and allow for archival appraisal at multiple decision points. The Archivematica micro-

services implement these requirements as granular system tasks which are provided by a combination of Python scripts and one or more of the free, open-source software tools bundled in the Archivematica system.

Archivematica uses METS, PREMIS, Dublin Core and other recognized metadata standards. The primary preservation strategy is to normalize files to preservation and access formats upon ingest when necessary (for example, when the file is in a format that is proprietary and/or is at risk of obsolescence). The media type preservation and access plans it applies during normalization are based on format policies derived from an analysis of the significant characteristics of file formats [4]. The choice of access formats is based on the ubiquity of viewers for the file format as well as the quality of conversion and compression. Archivematica's preservation formats are all open standards [5]. Additionally, the choice of preservation and access formats is based on community best practices and availability of open-source normalization tools.

Archivematica maintains the original files to support future migration and emulation strategies. However, its primary preservation strategy is to normalize files to preservation and access formats upon ingest. The default normalization format policies can be edited and disabled.

All of the software, documentation and development infrastructure are available free of charge and released under AGPL3 and Creative Commons licenses to give users the freedom to study, adapt and re-distribute these resources as best suits them. Archivematica development is led by Artefactual Systems, a Vancouver based technical service provider that works with archives and libraries to implement its open-source solutions as part of comprehensive digital preservation strategies. All funding for Archivematica development comes from clients that contract Artefactual's team of professional archivists and software developers to assist with installation, integration, training and feature enhancements. The majority of Archivematica users take advantage of its free and open-source license without additional contracting services.

3. ACQUISITION AND BACKLOG MANAGEMENT

Early implementers of the Archivematica suite of tools have consistently struggled with the mechanics of acquiring digital materials. Analogue records are delivered to the repository or are picked up from the donor's storage location, but digital acquisition can be more varied. Digital materials can arrive via digital transfer over a network such as email, FTP or shared directories. The archives may have to send an archivist to acquire the digital materials onsite, and even then, there are several options for acquisition including pickup, copying, or imaging. Depending on the type of acquisition, should the archivist photograph the condition of the materials in their original location? What steps must be taken to ensure that the digital objects copied or imaged retain their integrity during transfer to the archives? Finally, when digital materials are donated to the archives onsite, how do processes differ from pickup and digital network transfer?

Archivists who deal primarily with analogue materials are well accustomed to the need to maintain a backlog. Acquisitions regularly occur for which there are limited or no resources to process them immediately. For this reason, it is imperative that

the archives achieve a minimum level of control over the material so that it can be tracked, managed, prioritized and, if necessary, subjected to emergency preservation actions.

Archivematica runs through a set of transfer actions in the dashboard to establish initial control of the transfer. It verifies that the transfer is properly structured or structures it if necessary. Then, it assigns a unique universal identifier (UUID) for the transfer as a whole and both a UUID and a sha-256 checksum to each file in its /objects directory. Next, Archivematica generates a METS.xml document that captures the original order of the transfer and that will be included in any SIP(s) generated from this transfer. Any packaged files are unzipped or otherwise extracted, filenames are sanitized to remove any prohibited characters, and file formats are identified and validated. Finally, technical metadata is extracted from the files and the entire transfer content and metadata is indexed. At this point in the process, the transfer is ready to be sent to a backlog storage location that should be maintained in much the same way as the archival storage. The transfer is ready for future processing. These features will be added and evaluated in forthcoming releases of the Archivematica software.

4. ARRANGEMENT AND DESCRIPTION

Once an archive is ready to process one or more digital acquisitions, the next challenge comes from making a SIP from disparate parts of an acquisition. For example, in a situation in which an acquisition arrives on multiple digital media, the archives may have accessioned transfers from each media type and/or broken a very large hard drive into two or more transfers. Presumably, archivists will want their SIPs to be formed so that the resultant AIPs and DIPs conform to some level of their archival description, so SIP content could derive from one or more transfers or parts of transfers.

Arrangement and description do not neatly occur at one specific point during processing. Archivists arrange and describe analogue records intermittently. Arrangement is based upon the structure of the creator's recordkeeping system, inherent relationships that reveal themselves during processing and compensations made to simplify managing records and/or providing access. Archivists document their arrangement decisions and add this information, along with additional descriptive information gathered about the records during processing, to the archival description. Further, documentation of arrangement decisions and actions supports respect des fonds by preserving information about original order. Digital records must be arranged and described in order to effectively manage and provide access to them. Analogue functionality is very difficult to mimic in a digital preservation system such as Archivematica, because any interaction that allows for analysis of the records can result in changing original order and metadata associated with the records.

The OAIS model assumes that a digital archive system receives a fully formed SIP. However, this is often not the case in practice. Early Archivematica implementers were often manually compiling SIPs from transfers in the Thunar file browser bundled with the system. After transfer micro-services are completed successfully, Archivematica allows transfers to be arranged into one or more SIPs or for one SIP to be created from multiple transfers. The user can also re-organize and delete objects within the SIP(s). The original order of the transfer is maintained as its own structMap section in the transfer METS

file, a copy of which is automatically added to each SIP. Additionally, the archivist can use dashboard functionality to add basic descriptive metadata to the SIP at this point, including information about rights and restrictions.

The Archivematica project is now working on the ability to call up a transfer into a file browser interface in the dashboard's Ingest tab, examining its contents and forming it into SIPs for processing (See Figure 1).

review massive sets of digital records and compile selections from them as evidence. Clearly, the set of records presented as evidence must be verifiably authentic. Since archives are held to the same standards of authenticity there is much to be learned from the digital forensics field, which for over thirty years has been developing tools for processing evidence that guarantees its acceptance in courts. Such tools allow for auditing an investigator's actions, recording information about the set of

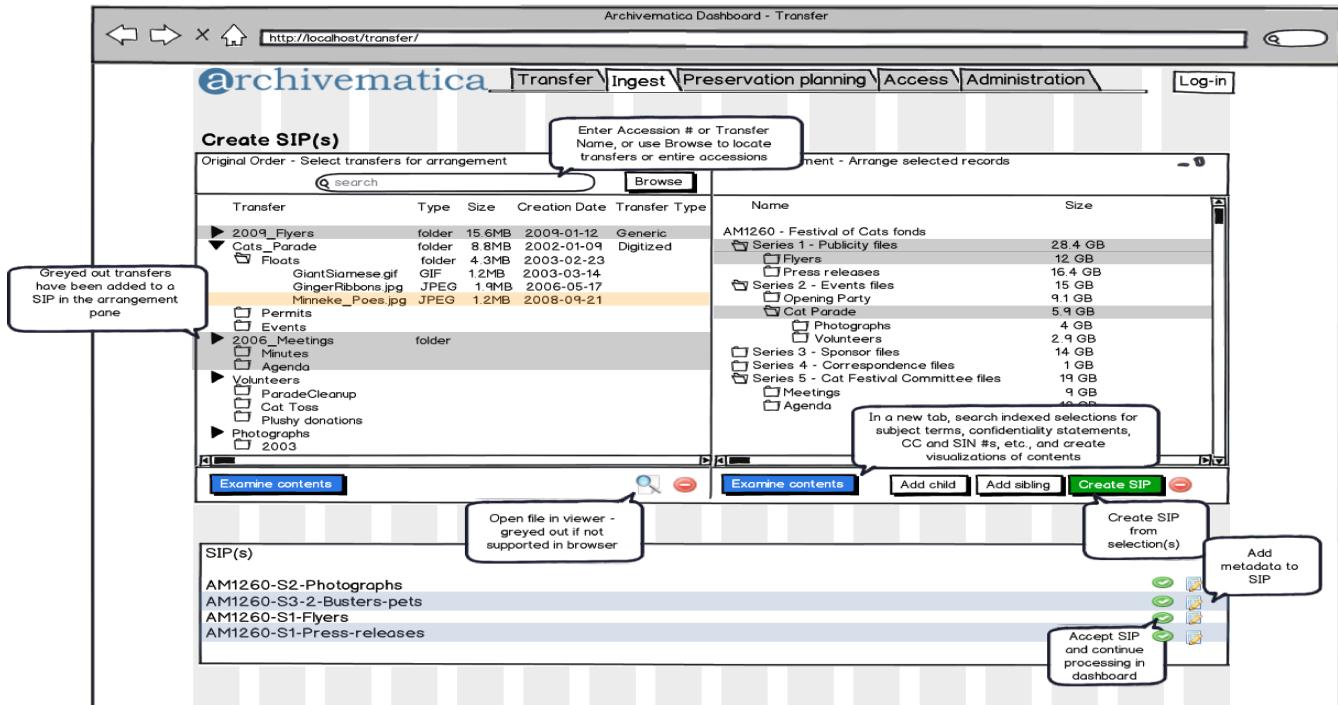


Figure 1. Create SIP dashboard interface mockup.

Much of the inspiration for such an interface came from digital forensics software and the Curator's Workbench at the University of North Carolina, Chapel Hill. The UNC Libraries had developed Curator's Workbench [6], a tool that, among other things, allows for arrangement of digital records without losing the original order. The Archivematica team considered including the tool in their suite, but because of concerns about integration and ongoing support, they opted instead to mimic its arrangement functionality. Archivematica's 0.8 alpha release uses the Xubuntu file browser Thunar to arrange records and METS to keep a record of the original order within each SIP formed from a transfer. In future releases, the METS file is still generated while file browser functionality has been moved to the dashboard. Future developments could see expanded METS and/or PREMIS profiles that includes information about selection actions undertaken during SIP creation and at the various appraisal stages.

The limitations for analyzing and forming SIPs using only the file browser were clear in earlier Archivematica releases. Transfers may contain restricted material, passwords, personal information or other content that is unsuitable for continued preservation. For insight into this problem, the project explored the possibilities of using digital forensics techniques. Digital forensics experts must

records and its origin while adding descriptive metadata and grouping portions of the set into discrete evidence packages, indexing and examining the file system structure and contents, and ensuring integrity. Many of the software tools used by digital forensics experts are proprietary, but in recent years open source tools have been developed to perform the same functions.

Despite their availability, open source digital forensics tools can be difficult to understand by non-experts. Serendipity's role in open source software development cannot be overstated. Just when Archivematica's systems analysts realized that they could not possibly decipher the entire canon of digital forensics software in time for the next release, digital humanities scholars and archivists in the United States were conceptualizing the BitCurator Project. From the BitCurator website [7]: "The BitCurator Project is an effort to build, test, and analyze systems and software for incorporating digital forensics methods into the workflows of a variety of collecting institutions." Artefactual Systems is closely involved with the BitCurator Project, with its president, Peter Van Garderen, on the Development Advisory Group and Courtney Mumma, systems analyst, on the Professional Experts Committee. Ideally, BitCurator will result in a set of open source tools that allow for arrangement,

description and other valuable functionalities that integrate well into the Archivematica suite.

Since open source digital forensics tools for archivists like BitCurator are not yet ready to be integrated into the Archivematica suite, the team looked for other ways to provide the necessary services to satisfy their workflows. One possible solution is using Apache Tika [8] and ElasticSearch [9] to index and search transfers in a dashboard file browser window to determine which part(s) to include in the SIP and to create visualizations of the transfer and SIP contents.

Requirements for future releases include indexing and reporting on all text content, file embedded metadata and file formats. Using Tika, ElasticSearch and other tools, Archivematica will provide keyword and pattern matching for privacy/security sensitive information (e.g. social insurance numbers/social security numbers, credit card numbers, email addresses and security keywords) and reports of such things as PDFs that have not been OCR'ed, password protected and encrypted files and duplicates with their full file paths. Reports for all of these indexing requirements will be available via the Examine Contents windows, accessible from the Create SIP browser window in the Ingest tab of the dashboard.

The Examine Contents reports will include a search box for the indexed transfer content, general information about the transfer or selected file group (e.g. number of files, size, name, UUID, and accession number), a pie graph visualization showing file type distribution overall and a bargraph visualization showing file type by folder and ordered by size. Clickable links will open to sub-reports on all contents of a specified format in context of the entire transfer, duplicates with their locations, privacy and confidentiality keywords and numbers and password protected files with their distribution across the entire contents visualized as a graph (See Figure 2).

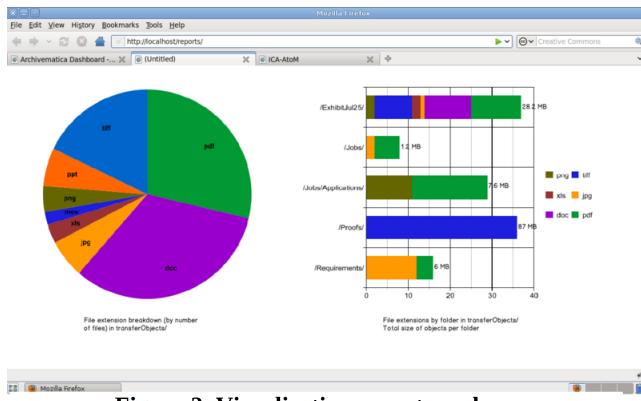


Figure 2. Visualization report mockup.

Such functions are being developed iteratively as part of the 2013 Archivematica 1.0 and subsequent releases. Should BitCurator or something else come along that can fulfill or expand on any of these functions, Archivematica's microservices architecture is such that the same requirements can be accomplished by these other tools with only minor changes to the code.

5. APPRAISAL

Originally intended for the long-term preservation of scientific data, OAIS does not address archival appraisal. To advise in the formation of appraisal requirements, the team consulted with the

InterPARES 3 Project [10] to conduct a gap analysis between OAIS and the InterPARES 1 Project's Chain of Preservation (COP) Model [11]. Review of the model, along with consultations with archivists about processing analogue records, revealed that appraisal occurs in a few different stages during archival processing. Archivists make an acquisition decision based on a preliminary appraisal, then reassess iteratively when they discover more about the records during accessioning actions and processing. Project partners including archivists and Archivematica developers built workflows around these different appraisal functions, which resulted in constructing three opportunities for appraisal in Archivematica: Selection for Acquisition, Selection for Submission and Selection for Preservation. The three appraisal opportunities as they were manifested in Archivematica 0.7.1 are discussed in detail in a recent Archivaria article [12], so the following is a brief summary of their functions and the associated Archivematica micro-services.

Selection for Acquisition occurs before records are accepted into an archives' custody for processing and preservation. Common practice in archives is to gather and review information about the records creator, the recordkeeping system(s) and the records to make an acquisition decision. For digital records, this includes learning as much as possible about the technological context of the records [8]. Because of limited access to originating technological environments for various reasons, it may become necessary for archives to acquire many more records than they might from an analogue body of records. Therefore, steps must be taken to ensure integrity of the records acquired while appraisal decisions are made over time.

Selection for Submission is the process of forming Submission Information Packages (SIPs) from acquired digital records or "transfers". In Archivematica, a transfer is any set of digital records acquired but not yet processed. Each SIP derives from one or more transfers. However, the SIP cannot be formed until the archivist has some information about the content of the transfer. For this reason, the transfer undergoes several micro-services first so that the archivist can review the results and assess how the received contents compare to the initial Selection for Acquisition expectations.

In the 0.8 alpha iteration of Archivematica, the archivist starts by adding a transfer to a specified folder in the file browser. The transfer begins processing in the Transfer tab of the web-based dashboard, where it is verified to be compliant for ingest in the system. Then, it is renamed with a transfer UUID and is assigned file UUIDs and checksums. If checksums already exist in the transfer, they are verified. A METS.xml file is added to the transfer, the transfer can be quarantined, and any packages are extracted. After a virus scan, prohibited characters are removed from filenames, formats identification process are run and metadata is characterized and extracted. All of the information generated from these micro-services allow the archivist to decide which parts of the transfer are archival materials ready for further processing.

In the 0.9 beta iteration of Archivematica, only one transfer can become one SIP, which is deprecated from the functionality of 0.8. In 0.8, one or more transfer(s) could become one or more SIP(s), but the arrangement was done in the file browser. The reason for this deprecation is so that the 1.0 release can move all the Selection for Acquisition functions to the web browser and

improve the tools to create SIPs as discussed in the Arrangement and Description section of this paper.

Selection for Preservation results in forming an Archival Information Package (AIP). A SIP is subjected to several micro-services, displayed in the Ingest tab, before the archivist has an opportunity to review the resulting AIP. Micro-services include verifying SIP compliance, renaming SIP with a SIP UUID, sanitizing file, directory and SIP name(s), checking integrity, copying metadata and logs from the transfer, and normalization. Once normalization and all other processing micro-services have run, the archivist can review the AIP contents and metadata in another browser window or download it to review using the file browser. At that point, they can either reject or accept the AIP and upload it into designated archival storage.

At every stage of appraisal, archivists may choose to destroy or deselect a record or set of records. Archivematica keeps logs of these changes by adding a text file listing excluded records to the logs directory in the transfer or SIP. This may even allow for richer and more transparent descriptive information about archival processing than is accomplished in analogue archives. It is important to note that the aforementioned steps are optional choices for the user. If the user has limited time or knows a great deal about the contents of a SIP, for instance, if the SIP is made up of described digitized videos, Archivematica can be configured to allow for automatic ingest.

In forthcoming releases, these appraisal processes will be incrementally moved to a web browser interface in the dashboard. Elastic Search indexing of the transfer and the AIP should also contribute to a richer, more informed selection process. Other development may include an automated process for “flagging” transfer content that may require further appraisal review based on a predefined set of indexing results.

6. PRESERVING AND PROVIDING ACCESS TO EMAIL

Several Archivematica project partners targeted email preservation as a priority in their digital archives planning. One pilot project involved acquiring a snapshot of the email account of a former university president. The account had been active for 10 years and no other email from the account had been sent to the university archives in electronic form in the past.

The university was using Zimbra Network Edition to send and receive email [13]. The Zimbra administrator's manual does not include information on how to export email from Zimbra for use in other email programs.[14] However, the university's IT department backs up the email accounts using a default directory structure specific to Zimbra, and was willing to deliver email to the Archives in the form of these backups. However, these backups are in a format which is intended to be used to restore email to Zimbra accounts, not to migrate the accounts' contents into other systems. Furthermore, documentation of its structure is somewhat limited. After analyzing the Zimbra backup and conducting research on email preservation standards and practices, the project team reached the conclusion that Zimbra email accounts need to be converted to a standard, well-documented, widely-used format that can be opened in a variety of open-source email programs or other tools such as web browsers.

Two formats which were explored as part of this project were Maildir and mbox [15]. Maildir is a text-based format which

stores each folder in an email account as a separate directory (inbox, sent items, subfolders etc) and each email as an individual text or .eml file [16]; attachments are included in the text files as base64 encoded ascii text. Mbox is a single large text file with attachments included as base64 content; each folder in an account is saved as a separate mbox file. Both formats can be imported into and rendered by numerous email programs, proprietary and open-source, and both can be converted into other formats using open-source tools and scripts. Although Maildir and mbox can be rendered in a variety of email programs, mbox has more potential as an access format because it is easier to develop tools to render it that are not necessarily email programs. For example, a program called Muse, developed by Stanford University [17], is designed to render mbox files using only a web browser. In addition, mbox is the source format for import into tools like the CERP email parser, which was developed by the Rockefeller Archive Center and the Smithsonian Institution Archives to convert email messages to hierarchically arranged XML files [18]. In essence, mbox is emerging as a de facto standard for which the digital curation community is beginning to build tools for rendering and manipulation. However, Maildir is preferable as a preservation format because it stores each message as a separate text file; thus any corruption to one or more text file would not cause an entire directory of messages to be lost, which is a risk with a format such as mbox.

The project team tested the use of a tool called OfflineImap [19] to back up a test Zimbra email account to Maildir and converted the Maildir backup to mbox using a freely available python script [20]. Following these preliminary tests, the Zimbra backup of the sample email account was restored to Zimbra and captured using OfflineImap. The resulting Maildir backup was converted to mbox files (Inbox, Sent and Eudora/out) which were imported into an open-source email program called Evolution. The total message count for each folder was found to be the same in Evolution as it had been in Zimbra (71, 2544 and 7628 messages, respectively), and randomly sampled emails were opened to ascertain that the conversion and import were successful. Sample emails from the Zimbra and Maildir backups were also compared to ensure that the significant characteristics of the Zimbra version were captured in the Maildir version [21].

A critical component of the University's email preservation strategy is management of access based on compliance with Freedom of Information and Protection of Privacy legislation. In any given user's account, some email messages must necessarily be excluded from public access based on the presence of personal information or other information which falls under exceptions to disclosure under the Act. The University's archivists and FOIPPA management personnel will need to be able to view email messages, flag those with restrictions, and provide public access to only those emails which are not restricted. Preliminary tests of Muse have shown it to be capable of importing mbox files, rendering the individual messages in a web browser, allowing tagging of restricted messages, and exporting the remainder in mbox format. We have noted that tagging one message as restricted automatically tags the same email message in other threads containing the same message.

Based on our analysis of pilot project email systems, email management practices, and preservation formats and conversion tools, we have summarized Archivematica requirements for acquiring, preserving and providing access to email. Ideally,

email is acquired, per account, in Maildir format, for the following reasons:

- The Maildir directory structure is well-documented and transparent;
- Maildir is widely used and can be created and rendered by a large number of software tools, both proprietary and open-source;
- OfflineIMAP is proving to be a useful tool for capturing email accounts in maildir format. Acting as an IMAP client, it can interact with a wide number of mail server programs, avoiding the need to add support for other mail server or email archive format conversions.
- The contents of a Maildir directory are plain text messages which can be read easily in any text editor (except for attachments);
- The text-based messages are based on an open and widely-used specification [22];
- Because each message is saved individually, accidental corruption or deletion of one or more messages would not result in the entire Maildir backup becoming unreadable (by comparison, corruption of a small amount of data in an mbox file could render the entire mbox file, with its multiple messages, unreadable);
- Maildir is easily converted to mbox for access purposes.

The archivists would submit the Maildir backup into Archivematica, where it would be retained as the preservation master in the AIP. Note that Maildir backups do not capture calendars or contact lists. However, University Archives staff have indicated that such records would probably not be considered archival. The attachments would be extracted and normalized to standard open formats for preservation purposes, with links between messages and their normalized attachments being managed through UIDs and/or filename. Attachments must be extracted and normalized because they pose a usability risk as base 64 ascii encoded text. They will always need to be rendered in a software program for human cognition of its content. In other words, even though the user may be able to open an email message in an email program he or she typically has to open the attachment separately using a software program that can render it.

For access, Archivematica will automatically generate a Dissemination Information Package (DIP) containing mbox files generated from the maildir preservation master. For an email account that consisted of an inbox with subfolders plus draft and sent items, the DIP would look something like this:

```
Inbox.mbox
Inbox.TravelCttee.mbox
Inbox.ExecCttee.mbox
Inbox.Workshops.mbox
Drafts.mbox
Sent.mbox
```

For most university and public repositories, provision of access must necessarily incorporate access and restriction management to comply with freedom of information, privacy and confidentiality requirements. The only known open-source tool that facilitates large-scale review and tagging of email account contents is Muse. More testing will be required to determine how usable and scalable the process of email tagging and exporting is with this tool. However, it should be noted that Muse is still in active development, and the Muse project team is interested in continuing to develop and refine the tool for use by libraries and archives. This bodes well for future feature development informed by Archivematica community members.

7. FORMAT POLICY REGISTRY - FPR

The Archivematica project team has recognized the need for a way to manage format conversion preservation plans, referred to by the project as format policies, which will change as formats and community standards evolve. A format policy indicates the actions, tools and settings to apply to a particular file format. The Format Policy Registry (FPR) will provide valuable online statistics about default format policy adoption as well as customizations amongst Archivematica users and will interface with other online registries (such as PRONOM and UDFR) to monitor and evaluate community-wide best practices. It will be hosted at archivematica.org/fpr.

An early prototype has been developed by Heather Bowden, then Carolina Digital Curation Doctoral Fellow at the School of Information and Library Science in the University of North Carolina at Chapel Hill (See Figure 3). A basic production version implementing these concepts will be included in upcoming releases. The FPR stores structured information about normalization format policies for preservation and access. These policies identify preferred preservation and access formats by media type. The choice of access formats is based on the ubiquity of viewers for the file format. Archivematica's preservation formats are all open standards; additionally, the choice of preservation format is based on community best practices, availability of open-source normalization tools, and an analysis of the significant characteristics for each media type. These default format policies can all be changed or enhanced by individual Archivematica implementers. Subscription to the FPR will allow the Archivematica project to notify users when new or updated preservation and access plans become available, allowing them to make better decisions about normalization and migration strategies for specific format types within their collections. It will also allow them to trigger migration processes as new tools and knowledge becomes available.

One of the other primary goals of the FPR is to aggregate empirical information about institutional format policies to better identify community best practices. The FPR will provide a practical, community-based approach to OAIS preservation and access planning, allowing the Archivematica community of users to monitor and evaluate formats policies as they are adopted, adapted and supplemented by real-world practitioners. The FPR APIs will be designed to share this information with the Archivematica user base as well with other interested communities and projects.

Extension	Normalization Description	Command Type	Command	Purpose
ac3	Transcoding to wav with ffmpeg	bashScript	ffmpeg -i "%file.FullName%" -ac:ffprobe -f wav "%file.FullName%.wav"	preservation
ac3	Transcoding to mp3 with ffmpeg	bashScript	ffmpeg -i "%file.FullName%" -ac:ffprobe -f mp3 "%file.FullName%.mp3"	access
af	Transcoding to wav with ffmpeg	bashScript	ffmpeg -i "%file.FullName%" -ac:ffprobe -f wav "%file.FullName%.wav"	preservation
af	Transcoding to mp3 with ffmpeg	bashScript	ffmpeg -i "%file.FullName%" -ac:ffprobe -f mp3 "%file.FullName%.mp3"	access

Figure 3 FPR format policies in early “Formatica” prototype. “Formatica” has since been renamed “FPR”.

8. CONCLUSION

Working with pilot project implementers, the Archivematica team has gathered requirements for managing a backlog of indexed digital acquisitions transfers, creating a SIP from a transfer or set of transfers, basic arrangement and description, preserving email, and receiving updates about new normalization paths via a format policy registry (FPR). After creating workflows that would account for real-world archival processing needs, these requirements have been added to our development roadmap for 0.9, 1.0 and subsequent Archivematica releases [23].

The Archivematica pilot project analysis and development described in this article are driven by practical demands from our early adopter community. The alpha release prototype testing sponsored by our contract clients and shared by a growing community of interested users from the archives and library professions and beyond has provided the opportunity to spearhead the ongoing evolution of digital preservation knowledge in the form of a software application that is filling a practical need for digital curators.

At the same time, the digital curation community is also evolving and maturing. New tools, concepts and approaches continue to emerge. The Archivematica technical architecture and project management philosophy are designed to take advantage of these advancements for the benefit of Archivematica users and the digital curation community at large.

The free and open-source, community-driven model provides the best avenue for institutions to pool their technology budgets and to attract external funding to continue to develop core application features as requirements evolve. This means the community pays only once to have features developed, either by in-house technical staff or by third-party contractors such as Artefactual Systems. The resulting analysis work and new software functionality can then be offered at no cost in perpetuity to the rest of the user community at-large in subsequent releases of the software. This stands in contrast to a development model driven by a commercial vendor, where institutions share their own expertise to painstakingly co-develop digital curation technology but then cannot share that technology with their colleagues or professional communities because of expensive and restrictive software licenses imposed by the vendor.

9. REFERENCES

- ISO 14721:2003, Space data and information transfer systems – Open archival information system – Reference model (2003).
- Artefactual Systems, Inc. and City of Vancouver, Requirements, <http://archivematica.org/wiki/index.php?title=Requirements> (accessed May 21, 2012).
- Artefactual Systems, Inc., Archivematica homepage, <http://archivematica.org> (accessed May 24, 2012).
- Archivematica significant characteristics evaluation, https://www.archivematica.org/wiki/Significant_characteristics (accessed August 19, 2012).
- Wikipedia definition of open standards, http://en.wikipedia.org/wiki/Open_standard (accessed August 17, 2012).
- Carolina Digital Repository Blog, “Announcing the Curator’s Workbench”, <http://www.lib.unc.edu/blogs/cdr/index.php/2010/12/01/announcing-the-curators-workbench/> (accessed May 21, 2012).
- BitCurator Tools for Digital Forensics Methods and Workflows in Real-World Collecting Institutions, <http://www.bitcurator.net/> (accessed May 21, 2012).
- Tika website, <http://tika.apache.org/> (accessed May 21, 2012).
- ElasticSearch website, <http://www.elasticsearch.org/> (accessed May 21, 2012).
- .InterPARES 3 Project, http://www.interpares.org/ip3/ip3_index.cfm (accessed May 21, 2012).
- .InterPARES 2 Project, Chain of Preservation (COP) Model, http://www.interpares.org/ip2/ip2_model_display.cfm?model=cop (accessed May 21, 2012).
- Courtney C. Mumma, Glenn Dingwall and Sue Bigelow, “A First Look at the Acquisition and Appraisal of the 2010 Olympic and Paralympic Winter Games Fonds: or, SELECT * FROM VANOC_Records AS Archives WHERE Value='true';” (Archivaria 72, Fall 2011) pgs. 93-122.
- Zimbra website, <http://www.zimbra.com/> (accessed May 21, 2012).
- Administration Guide to Zimbra, http://www.zimbra.com/docs/me/6.0.10/administration_guide/ (accessed May 24, 2012).
- Wikipedia articles describing maildir and mbox., http://en.wikipedia.org/wiki/Maildir_and_mbox. Note that this paper refers specifically to the .mbox extension, the standard Berkeley mbox implementation of this format. For another discussion of the role of mbox in email preservation, see Christopher J. Prom, “Preserving Email,” DPC Technology Watch Report 11-01 (December 2011), <http://dx.doi.org/10.7207/twr11-01>. (accessed May 23, 2012).

16. EML is a common email format encoded to the RFC 822 Internet Message Format standard (<http://tools.ietf.org/html/rfc822>) for individual emails. Messages in Maildir backups are encoded to this standard, although they lack the .eml file extension. For a discussion of the role in the eml format in email preservation, see Prom, “Preserving email”.
17. Muse website, <http://mobilisocial.stanford.edu/muse/> (accessed May 21, 2012).
18. CERP XML format, <http://siarchives.si.edu/cerp/parserdownload.htm>. The CERP XML format is designed to be a neutral, software-independent format for email preservation, but as yet there are no tools available to display the XML files as email messages that can easily be searched and navigated.
19. Offline Imap website, <http://offlineimap.org/> According to the documentation for this tool, it is possible to specify the folders to be captured, which would permit capturing folders designated specifically for archival retention. OfflineImap can also be run as a cron job, capturing email automatically at specified intervals. These features open up a number of possibilities for email archiving workflows.
20. Python script, md2mb.py, available from <https://gist.github.com/1709069>.
21. Significant characteristics analysis for maildir, http://www.archivematica.org/wiki/index.php?title=Zimbra_to_Maildir_using_OfflineImap for an example of the analysis of significant characteristics. (accessed May 24, 2012).
22. RFC # 822, Standard for the Format of ARPA Internet Text Messages, <http://tools.ietf.org/html/rfc822>.
23. Archivematica Development Roadmap, https://www.archivematica.org/wiki/Development_roadmap/ (accessed August 21, 2012).

Preserving Electronic Theses and Dissertations: Findings of the *Lifecycle Management for ETDs* Project

Martin Halbert
University of North Texas
1155 Union Circle #305190
Denton, TX, 76203
940-565-3025
martin.halbert@unt.edu

Katherine Skinner
Educopia Institute
1230 Peachtree Street
Atlanta, GA 30309
404-783-2534
katherine@metaarchive.org

Matt Schultz
MetaArchive Cooperative
1230 Peachtree Street
Atlanta, GA 30309
616-566-3204
matt.schultz@metaarchive.org

ABSTRACT

This paper conveys findings from four years of research conducted by the MetaArchive Cooperative, the Networked Digital Library of Theses and Dissertations (NDLTD), and the University of North Texas to investigate and document how academic institutions may best ensure that the electronic theses and dissertations they acquire from students today will be available to future researchers.

Categories and Subject Descriptors

E.1 [Data Structures]: *distributed data structures*. H.3.2 [Digital Libraries]: *Information Storage, file organization*. H.3.4 [Systems and Software]: *distributed systems*. H.3.6 [Library Automation]: *large text archives*. H.3.7 [Digital Libraries]: *collection, dissemination, standards, systems issues*.

General Terms

Management, Documentation, Performance, Design, Reliability, Standardization, Languages, Theory, Legal Aspects, Verification.

Keywords

Archival Information Packages, Data Management, Digital Archives, Digital Curation, Digital Libraries, Electronic Theses and Dissertations, ETDs, Digital Objects, Digital Preservation, Distributed Digital Preservation, Ingest, Interoperability, Micro-Services, Repository Software, Submission Information Packages.

1. INTRODUCTION

One of the most important emerging responsibilities for academic libraries is curatorial responsibility for electronic theses and dissertations (ETDs) which serve as the final research products created by new scholars to demonstrate their scholarly competence. These are important intellectual assets both to colleges and universities and their graduates. Because virtually all theses and dissertations are now created as digital products with new preservation and access characteristics, a movement toward ETD curation programs in both U.S. institutions and abroad began in the early 1990's and has continued to this day.

There are many articles documenting this movement. The Coalition for Networked Information (CNI) recently studied the history of ETDs and graduate education and conducted an international survey concerning ETDs that examined linkages between the growth of ETD programs, institutional repositories, open access and other important trends in higher education (Lippincott and Lynch, 2010). Additional key issues identified in

the CNI survey are questions and uncertainty within institutions concerning ETD embargoes, ETD format considerations, costs of ETD programs, and the role of libraries in working with graduate schools to maximize benefits of ETD programs for students.

A basic point made by the CNI study and virtually all current literature on the ETD movement is that colleges and universities have been steadily transitioning from traditional paper/microfilm to digital submission, dissemination, and preservation processes. Increasingly, academic institutions worldwide are now accepting and archiving *only* electronic versions of their students' theses and dissertations, especially in archiving programs operated by academic libraries. While this steady transition in curatorial practice from print to digital theses and dissertations greatly enhances the current accessibility and sharing of graduate student research, it also raises grave long-term concerns about the potential ephemerality of these digital resources.

Our research focuses on answering the question: *How will institutions address the entire lifecycle of ETDs, ensuring that the electronic theses and dissertations they acquire from students today will be available to future researchers?* We use the phrase *lifecycle management of digital data* in the broad sense defined by the Library of Congress to refer to the "progressive technology and workflow requirements needed to ensure long-term sustainability of and accessibility to digital objects and/or metadata" (Library of Congress, 2006), as well as in the more detailed senses of the digital lifecycle management model as articulated by the Digital Curation Centre in the UK (Higgins, 2008). A key outcome of our research and documentation will be a clearly articulated lifecycle model specific for ETDs.

In order to unpack this complex issue and to assess the library field's ETD lifecycle-management needs and practices, leaders of the Networked Digital Library of Theses and Dissertations (NDLTD) and the MetaArchive Cooperative conducted a series of investigations during 2008-2010. These efforts included surveys, a pilot project, and meetings of the leadership of the two groups, each of which are concerned with different aspects of preserving ETDs. The research team then embarked upon a US Institute for Museum and Library Services-funded project in 2011 to develop guidelines for ETD lifecycle management, software tools to facilitate ETD curation, and educational materials to help prepare ETD curators. As one component of this project, we conducted a focus group with stakeholders. We describe our findings from these surveys below.

1.1 Surveys of ETD Curation Practices

In order to assess practitioner needs and the current status of the field, the MetaArchive Cooperative and the NDLTD conducted a survey in 2007/2008 to examine ETD practices and associated concerns in institutions either currently engaged in ETD programs or considering such preservation service programs. The on-line survey was distributed through five major listservs and received 96 responses, primarily from academic institutions that were providing or strongly considering collection of ETDs and associated ETD services (McMillan, 2008).

Of the survey respondents, 80% accept ETDs, and 40% accept *only* ETDs. The ETD programs report that they accept many formats (more than 20) beyond PDF documents, including images (92%), applications (89%), audio (79%), text (64%) and video (52%). The average size of these programs was 41 GB, and respondents reported 4.5 GB/year average growth. We found that the repository structures used by respondents also vary widely. The more popular approaches included locally developed solutions (34%), DSpace (31%), ETD-db (15%), and such vendor-based repositories as bepress (6%), DigiTool (6%), ProQuest (6%), and CONTENTdm (6%).

This diversity of system types—presumably at least somewhat representative of the overall industry—presents an array of challenges for preservation. Each of these repository systems requires preservation attention during the ingest process to ensure that the materials are submitted in such a way that it is possible to retrieve them and repopulate that repository system with the content. This demands that content carries with it a level of context, and that context differs across repository structures.

The digital collections file and folder structures used by respondents also varied widely. Most respondents reported that their ETD collections are not structured in logically named, manageable virtual clusters. In fact, more than a quarter of respondents reported that their ETD collections are stored in one mass upload directory. This raises many preservation readiness challenges. How can the institution preserve a moving, constantly growing target? How can they ensure that embargoed and non-embargoed materials that often co-exist in the same folder are dealt with appropriately? How will the institution know what these files are if they need to repopulate their repository with them, particularly if they are stored in a repository system that does not elegantly package metadata context with content at export? Only 26% of the institutions manage their ETD collections in annual units. Another 26% use names (departments, authors) or disciplines as unit labels. Seven percent reported using access level labels and another 13% did not know.

The survey also collected information about what information institutions would need to make decisions concerning ETD preservation programs. Perhaps the most remarkable finding from this survey was *that 72% of responding institutions reported that they had no preservation plan for the ETDs they were collecting.*

The responses to this survey led the same researchers to conduct a follow-on survey in 2009 that probed more deeply into digital preservation practices and concerns (Skinner and McMillan, 2009). This survey included questions concerning institutional policies, knowledge and skills needed for digital preservation activities, level of desire for external guidance and expertise in digital preservation, and perceptions about relative threat levels of different factors in the long-term survivability of digital content.

Based on these findings, the MetaArchive Cooperative and the NDLTD undertook a joint pilot project in 2008-2010 to further explore and understand issues highlighted in the surveys and to respond to concerns of their respective memberships about preservation of ETDs. In the course of this pilot project, a group of institutions that are members of both organizations (including Virginia Tech, Rice University, Boston College, and others) worked together to discuss, analyze, and undertake experiments in different aspects of lifecycle management of ETDs, and to identify problem areas experienced by multiple institutions. The pilot project group also explored the literature to better understand what has been published to date on different digital lifecycle management topics, and how such publications relate to ETDs.

During this pilot project, as another means of assessing needs, Gail McMillan (NDLTD) and Martin Halbert (MetaArchive Cooperative) asked a large number of ETD program leaders about their concerns about ETD lifecycle management during workshops conducted at each of three annual ETD conferences hosted by the NDLTD from 2008-2010. Findings from the pilot project analysis and workshop inquiries were reviewed and discussed at three joint planning meetings of the NDLTD board and MetaArchive leadership during this period. They were consistent with the initial findings of the 2007-8 ETD survey.

Similarly, as the *Lifecycle Management for ETDs* project kicked off in 2012, the research team hosted a focus group in conjunction with the February Texas Electronic Theses and Dissertations Association meeting in Denton, Texas. Respondents in this focus group included both College of Arts and Sciences representatives and library representatives. The concerns raised by this group mirrored our earlier findings—most are involved in ETD programs and are either already electronic *only* or will be in the near future. The collection structures, file-types accepted, and repository infrastructures vary wildly. All attendees agreed that establishing documentation, tools, and educational materials that encourage better, more consistent ETD curatorial practices are of great need and should be of value to virtually all categories of academic institutions within the United States and internationally.

2. GUIDANCE DOCUMENTS

There is need for guidance documents in a variety of specific ETD lifecycle management topics to advance the capabilities of institutions that administer ETD service programs. The *Lifecycle Management for ETDs* project has worked to fill these gaps. The research team strongly feels that as a field we need to better understand, document, and address the challenges presented in managing the entire lifecycle of ETDs in order to ensure that colleges and universities have the requisite knowledge to properly curate these new collections. The research team has developed draft documentation on a number of topical areas, as briefly described below.

2.1 Introduction to ETDs

Prepared by Dr. Katherine Skinner and Matt Schultz (Educopia, MetaArchive), this document introduces the “Guidelines” and chronicles the history of ETDs. Using survey data and research findings, it describes the evolving and maturing set of practices in this area. It discusses the philosophical and political issues that arise in this genre of content, including what to do with digitized vs. born-digital objects, how to make decisions about outsourcing, and how to deal with concerns about future publications and

embargoed materials in the lifecycle management framework. The chapter provides a conceptual overview of a lifecycle model for ETDs that makes direct connections between the model and the individual guidance documents described below.

2.2 Access Levels and Embargoes

Prepared by Geneva Henry (Rice University), this document provides information about the ramifications of campus policy decisions for or against different kinds of access restrictions. It defines access restriction and embargo, and discusses reasons for each, including publishing concerns, sensitivity of data, research sponsor restrictions, and patent concerns. It discusses how institutions may provide consistent policies in this area and how policies might impact an institution's lifecycle management practices. It also reviews and compares existing university policies and makes policy recommendations.

2.3 Copyright Issues and Fair Use

Patricia Hswe (Penn State) chronicles ETD copyright and fair use issues that arise both in the retrospective digitization and the born-digital acquisition of theses and dissertations. It discusses institutional stances and guidelines for sponsored research and student work, and also reviews copyright and fair use issues with respect to commercial publishers (including e-book publishers) and vendors such as ProQuest. It seeks to provide clarifying information concerning publisher concerns and issues, providing a concise summary of the relevant information for stakeholders.

2.4 Implementation: Roles & Responsibilities

Xiaocan (Lucy) Wang (Indiana State University) documents the variety of stakeholders who impact and are impacted by the transition to electronic submission, access, and preservation of theses and dissertations, including such internal stakeholders as institutional administration (e.g., president, provost, CIO, general counsel), graduate schools (administrators, students, faculty), libraries (administrators, digital initiatives/systems divisions, technical services, reference), and offices of information technology, and such external stakeholders as commercial vendors/publishers, NDLTD, access harvesters (e.g., OCLC), and digital preservation service providers (e.g., MetaArchive, FCLA, DuraCloud). It emphasizes the range of functions played by these stakeholders in different management phases and institutions.

2.5 Demonstrations of Value

Dr. Yan Han (University of Arizona) provides guidance for institutions concerning assessment of ETD usage, and how communicating such assessment metrics can demonstrate a program's benefits to stakeholders. Han also documents practical examples of documenting and conveying usage metrics for stakeholder audiences, including the university, the students, and the research community more generally. He provides practical guidance for collecting, evaluating, and interpreting usage metrics in support of ETD programs, and discusses how it may be used to refine and promote this collections area.

2.6 Formats and Migration Scenarios

What factors should be considered by colleges and universities to determine what formats they should accept? How can they manage on an ongoing basis the increasingly complex ETDs that are now being produced by students? Bill Donovan (Boston College) discusses these format issues, including "data wrangling" practices for legacy content and migration scenarios for simple and complex digital objects in ETD collections.

2.7 PREMIS Metadata and Lifecycle Events

Another issue revealed in the needs assessment process was that most institutions do not have workflows and systems in place to capture the appropriate levels of metadata needed to manage ETDs over their entire lifecycle. Daniel Alemneh (University of North Texas) informs stakeholders and decision makers about the critical issues to be aware of in gathering and maintaining preservation metadata for ETDs, not just at the point of ingestion, but subsequently, as ETDs often have transitional events in their lifecycle (embargo releases, redactions, etc.). This guidance document will both inform and reinforce the software tools around PREMIS metadata that we are building.

2.8 Cost Estimation and Planning

Gail McMillan (Virginia Tech) provides institutions with information on costs and planning, laying out the critical paths that many ETD programs have charted to date. This document provides cost-benefit analyses of multiple scenarios to give institutions a range of options to consider for their local needs.

2.9 Options for ETD Programs

Our surveys and focus group have demonstrated that many institutions are delayed in ETD program planning simply because they do not have a clear understanding of the range of options to consider in implementing an ETD program. Restricted or open access? Implement an ETD repository or lease a commercial service? Who has responsibility for what functions? Dr. Martin Halbert (University of North Texas) explains the relevant decisions institutions must make as they set up an ETD program and clarifies the pros and cons of different options.

3. LIFECYCLE MANAGEMENT TOOLS

The research team is developing and openly disseminating a set of software tools to address specific needs in managing ETDs throughout their lifecycle. These tools are modular micro-services, i.e. single function standalone services that can be used alone or incorporated into larger repository systems. Micro-services for digital curation functions are a relatively new approach to system integration pioneered by the California Digital Library and the Library of Congress, and subsequently adopted by the University of North Texas, Chronopolis, MetaArchive, Archivematica, and other digital preservation repositories.

The micro-services described below draw upon other existing open source software tools to accomplish their aims. The intent of creating these four micro-services is that they will catalytically enhance existing repository systems being used for ETDs, which often lack simple mechanisms for these functions.

3.1 ETD Format Recognition Service

Accurate identification of ETD component format types is an important step in the ingestion process, especially as ETDs become more complex. This micro-service will: 1) Enable batch identification of ETD files through integration of function calls from the JHOVE2 and DROID format identification toolkits; and 2) Structure micro-service output in ad hoc tabular formats for importation into repository systems used for ETDs such as DSpace, and the ETD-db software, as well preservation repository software such as iRODS and DAITSS and preservation network software such as LOCKSS.

Components & Basic Requirements:

JHOVE2, DROID, XML output schema, Utility scripts (run commands, output parsers, etc.) & code libraries, API function calls, System requirements, Documentation & instructions

3.2 PREMIS Metadata Event Record-keeping

One gap highlighted in the needs analysis was the lack of simple PREMIS metadata and event record keeping tools for ETDs. This micro-service needs to: 1) Generate PREMIS Event semantic units to track a set of transitions in the lifecycle of particular ETDs using parameter calls to the micro-service; and 2) Provide profile conformance options and documentation on how to use the metadata in different ETD repository systems.

Components & Basic Requirements:

PREMIS Event profiles (example records) for ETDs, Event-type identifier schemes and authority control, AtomPub service document & feed elements, Utility scripts (modules) & code libraries, API function calls, Simple database schema & config, System requirements, Documentation

3.3 Virus Checking

Virus checking is an obvious service needed in ETD programs, as students' work is often infected unintentionally with computer viruses. This micro-service will: 1) Provide the capability to check ETD component files using the ClamAV open source email gateway virus checking software; 2) Record results of scans using the PREMIS metadata event tracking service; and 3) Be designed such that other anti-virus tools can be called with it.

Components & Basic Requirements:

ClamAV, Utility scripts (run commands, output parser, etc.) & code libraries, API function calls, System requirements, Documentation & instructions

3.4 Digital Drop Box with Metadata Submission Functionality

This micro-service addresses a frequently sought function to provide a simple capability for users to deposit ETDs into a remote location via a webform that gathers requisite submission information requested by the ETD program. The submission information will: 1) Generate PREMIS metadata for the ETD files deposited; 2) Have the capacity to replicate the deposited content securely upon ingest into additional locations by calling other Unix tools such as rsync; and 3) Record this replication in the PREMIS metadata.

Components & Basic Requirements:

Metadata submission profile(s), Client/server architecture, GUI interface, SSL, authentication support, Versioning support, Various executables, scripts & code libraries, Database schema & config, System requirements, Documentation

All of these tools will be documented and released in 2013 via the project site: <http://metaarchive.org/imls>.

4. CONCLUSIONS

The first phase of this project has helped to reinforce preliminary research we had conducted regarding ETD lifecycle management practices (or the significant lack thereof). The field has a dire need

for descriptive, not prescriptive, documentation regarding the range of ETD programs that institutions have designed and implemented to date, and the variety of philosophical, organizational, technical, and legal issues that are embedded therein. The field also has a stated need for lightweight tools that can be quickly implemented in a range of production environments to assist with some of the commonly needed curatorial practices for lifecycle management of these collections.

5. ACKNOWLEDGMENTS

We greatly appreciate the generous support of the Institute for Museum and Library Services (IMLS).

6. REFERENCES

- Caplan, Priscilla. "The Preservation of Digital Materials." *Library Technology Reports*, (2008) 44, no. 2.
- Conway, Paul. "Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas." *Library Quarterly*, (2010) 80:1, 61-79.
- Fox, Edward A., Shahrooz Feizabadi, Joseph M. Moxley, and Christian R. Weisser, eds. *Electronic Theses and Dissertations: A Sourcebook for Educators, Students, and Librarians*. New York: Marcel Dekker, 2004.
- Halbert, Martin, Katherine Skinner and Gail McMillan. "Avoiding the Calf-Path: Digital Preservation Readiness for Growing Collections and Distributed Preservation Networks," *Archiving 2009*, Arlington, VA, May 2009, p. 86-91.
- Halbert, Martin, Katherine Skinner and Gail McMillan. "Getting ETDs off the Calf-Path" ETD 2009: *Bridging the Knowledge Divide*, Pittsburgh, PA, June 10-13, 2009. Sharon Reeves, ed. <http://conferences.library.pitt.edu/ocs/viewabstract.php?id=733&cf=7>
- Hall, Susan L., Lona Hoover, and Robert E. Wolverton, Jr.. "Administration of Electronic Theses/Dissertations Programs: A Survey of U.S. Institutions." *Technical Services Quarterly* 22, no. 3 (2005): 1-17.
- Lippincott, Joan K. "Institutional Strategies and Policies for Electronic Theses and Dissertations." *EDUCAUSE Center for Applied Research Bulletin*, no. 13 (2006). <http://net.educause.edu/ir/library/pdf/ERB0613.pdf>
- Lippincott, Joan K., and Clifford A. Lynch. "ETDs and Graduate Education: Programs and Prospects." *Research Library Issues*, no. 270 (June 2010): 6-15. <http://publications.arl.org/rli270/>
- McMillan, Gail. "ETD Preservation Survey Results." *Proceedings of the 11th International Symposium on ETDs*, Robert Gordon University, Aberdeen, Scotland. (June 2008) <http://scholar.lib.vt.edu/staff/gailmac/ETDs2008PreservPaper.pdf>
- McMillan, Gail, and Katherine Skinner. (2010) "NDLTD/MetaArchive Preservation Strategy." (3rd ed.) <http://scholar.lib.vt.edu/theses/preservation/NDLTDPreservationPlan2010.pdf>
- Skinner, Katherine, and Gail McMillan. "Surveys of Digital Preservation Practices and Priorities in Cultural Memory Organizations." 2009 NDIIPP Partners Meeting, Washington, D.C., June 24, 2009. http://www.digitalpreservation.gov/news/events/ndiipp_meetings_ndiipp09/docs/June24/NDIIPP_Partners_2009_finalRev2.ppt

Creating Visualizations of Digital Collections with Viewshare

Trevor Owens

Library of Congress

101 Independence Ave. S.E.

Washington, D.C., 20540, U.S.A.

trow@loc.gov

Abigail Potter

Library of Congress

101 Independence Ave. S.E.

Washington, D.C., 20540, U.S.A.

abpo@loc.gov

ABSTRACT

Viewshare is a free, Library-of-Congress-sponsored platform that empowers historians, librarians, archivists and curators to create and customize dynamic interfaces to collections of digital content. This demonstration of Viewshare will start with an example spreadsheet or data harvested via OAI-PMH to generate distinct interactive visual interfaces (including maps, timelines, and sophisticated faceted navigation), which can be copy-pasted in any webpage. The data augmentation services associated with Viewshare will also be demonstrated.

Categories and Subject Descriptors

D.0 [General]: Software

General Terms

Design, Experimentation.

Keywords

Access, metadata, visualization.

1. DEMONSTRATION

Digital cultural heritage collections include temporal, locative, and categorical information that can be tapped to build interfaces to build dynamic interfaces to these collections. These kinds of dynamic interfaces are increasingly the way end users expect to interact with online content. However, they are often expensive and time consuming to produce

Simply put, search is not enough. End users want to browse content on a map, interact with it on a timeline, and dynamically pivot through data on the screen. The Viewshare project was created to make it as easy as possible for anyone working with cultural heritage collections to create these interfaces.

Briefly, Viewshare is a free platform built by Zepheira LLC for the Library of Congress which empowers historians, librarians, archivists and curators to create and customize views, (interactive maps, timelines, facets, tag clouds) of digital collections which allow users to interact with them in intuitive ways. The demonstration will cover how users can use the software to ingest

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

iPres '12, October 1-5, 2012, Toronto, Ontario, Canada.

Copyright 2010 ACM 1-58113-000-0/00/0010 . . . \$15.00.

collections from spreadsheets or MODS records, augment and transform their data online, generate distinct interactive visual interfaces, (including maps and timelines, and sophisticated faceted navigation) and ultimately copy-paste to embed the interfaces they design in any webpage.

The use of Viewshare does not require any specific technical skills or software. Any individual associated with a cultural or historical organization is encouraged to sign up for an account at <http://viewshare.org>.

2. THE VIEWSHARE WORKFLOW

Users import and augment existing collection data, iteratively build interfaces to their collection data and ultimately are able to share the interfaces and views which they have created. Viewshare interfaces are built entirely upon user-uploaded metadata. Recognizing the heterogeneity of collection data, Viewshare allows multiple methods of importing data. Users can build or work from existing simple spreadsheets, MODS records, and import Dublin Core metadata via OAI-PMH. To make this data usable, Viewshare includes a set of data augmentation tools to work from this extent data. For example, Viewshare enables users to derive latitude-longitude coordinates from plain text place names and then use these coordinates to plot their items on a map. Similarly, plain text expressions of date information can be used to derive ISO 8601 formatted dates for plotting items on a timeline. With its ease-of-ingest and data augmentation features, Viewshare understands and facilitates the use of the unique and sometimes idiosyncratic nature of cultural heritage collection metadata. At the same time, it also allows users to enhance this metadata in order to power the creation of dynamic interfaces.

After importing and augmenting collection data users begin creating interfaces. The tool's primary purpose is building dynamic, interactive views of digital collections. Through a drag-and-drop interface, users can create multiple views including maps, timelines, charts, and other dynamic visualizations. Users can then choose which facets they want to include in order to create unique ways of manipulating the data presented in each of the views. For instance, in a collection of postcards, a tag cloud facet set to display subject information will show the relative frequency of the subjects throughout the collection. If a user clicks on one of those subjects, Viewshare will limit the display of whatever view they are using to show only the objects associated with that term. As a user selects the data values they want to use in a given facet, and the particular views they want to display, they can use the "show preview" function to continually toggle back and forth between building their interface and a fully functional preview of what their resulting interface will look like. In this way, the tool supports an iterative and exploratory approach to creating these interfaces.

3. A VIEWSHARE EXAMPLE

After uploading a spreadsheet of the collection data, which includes links to the web-accessible image files, a user can begin building new interactive views. The original collection data includes plain-text place names which Viewshare can convert to points of latitude and longitude. With that data, a user can add a map showing the exact location of each card's creator. A clickable pin on the map allows users to see a thumbnail image of the item and select metadata elements. By adding a facet to the view, a

user can click on any facet element, such as subject heading "flowers," and the map will update to show only the location of the flower trade cards. Adding other facets such as date or business type will allow a user to further manipulate the geographic display. Additional interfaces, such as timelines, charts, galleries, tables, and other visualizations can be created—all with the same faceting, sliders, and discovery elements.

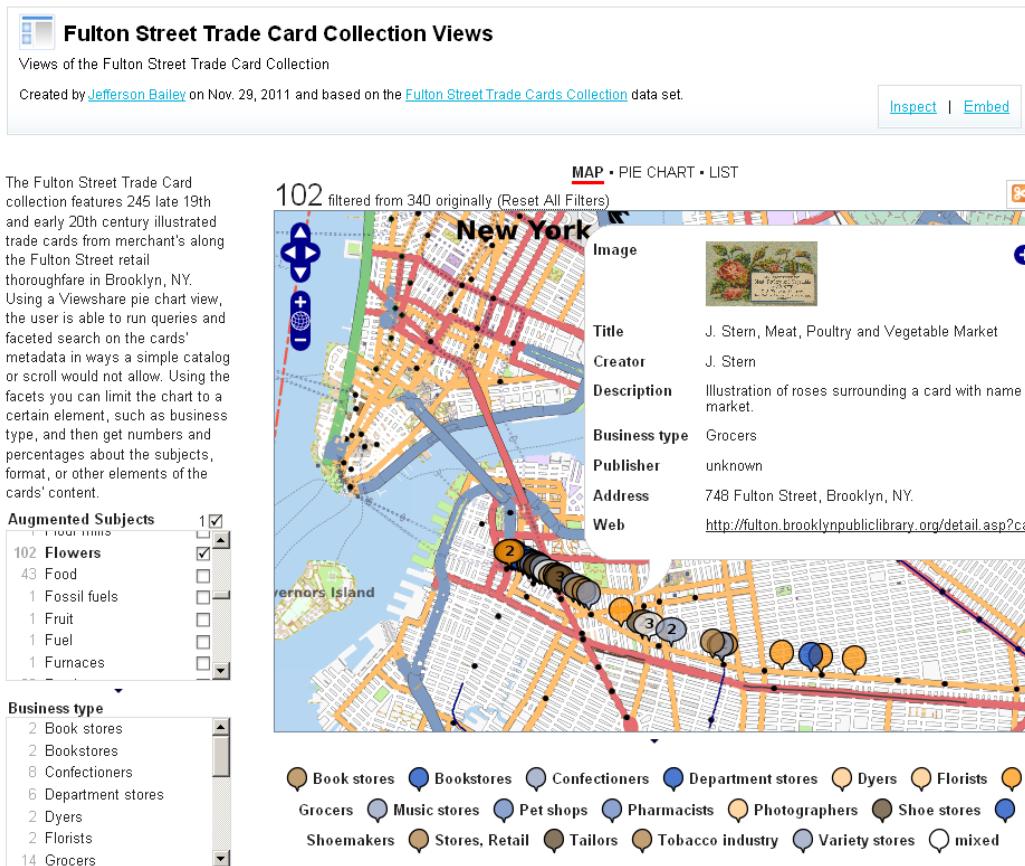


Figure 1. Screenshot of Fulton Street Trade Card Collection View: The user has selected to facet their display to only show cards with the subject "flowers" and has clicked on one of the orange pins associated with a grocer business type. As a result, Viewshare is now displaying a particular grocer's card associated with that address.

At the heart of building views is the ability to toggle between the "build" screen and the "preview" screen. Creating visualizations using different data elements from the collection offers an iterative, exploratory way to discover new relations between items, to excavate new meanings and promote new ways of understanding digital material. This back-and-forth modality characterizes many of Viewshare's features as well as its conceptual goals. Iterative interface construction encourages both close and distant readings; it empowers both the deep knowledge of collection stewards and the unguided explorations of regular users; it provides tools for both curatorial intent and algorithmic serendipity; and it encourages access, sharing, and linked open data.

2. INTENDED IMPACT

Curators of digital collections will benefit from this demonstration. They will see how easy it is to use Viewshare to produce interactive interfaces and enhance access to digital collections. Curators without access to web designers will especially benefit because they will be able to create tools like maps, faceted browsing and timelines—tools that are increasingly becoming the standard way of exploring content on the web—by themselves.

3. ACKNOWLEDGEMENTS

Viewshare is an open source project developed by Zepheira LLC. Source code is available for download at: <http://www.sourceforge.net/projects/loc-recollc>

Scalable content profiling for preservation analysis

Petar Petrov
Vienna University of Technology
Vienna, Austria
petrov@ifs.tuwien.ac.at

Christoph Becker
Vienna University of Technology
Vienna, Austria
becker@ifs.tuwien.ac.at

ABSTRACT

The starting point of any operational endeavor to preserve digital content is gaining a deep understanding of the characteristics of the objects. Systematic analysis of digital object sets and the identification of sample objects that are representative of a collection are critical steps towards preservation operations and a fundamental enabler for successful preservation planning. Without a full understanding of the properties and peculiarities of the content at hand, informed decisions and effective actions cannot be taken. This article presents a software tool prototype that is able to profile large sets of meta data in a scalable fashion and provide deeper insight into the digital collection at hand.

Categories and Subject Descriptors

H.3.1 [Information Systems]: Information Storage and Retrieval—*Content Analysis and Indexing*; H.3.7 [Information Systems]: Digital Libraries—*Collection*

Keywords

Digital Preservation, Preservation Planning, Content Profiling, Characterization, Stratification, Scalability

1. INTRODUCTION

Digital preservation is increasingly becoming relevant for large-scale collections, due to the increase of born-digital material in recent years. Content holding institutions commonly deploy digital content repositories that provide content management facilities and support for large data volumes. However, there is often no comprehensive overview about the detailed types of data contained. Apart from general information such as the number of objects, formats, mime-types and size, there is often a lack of deeper knowledge about the digital objects at hand. Although the meta data of each object can be produced automatically, currently there is no easy way of obtaining a deeper insight into digital collections. The starting point of digital preservation is the deep understanding of the objects at risk. Preserva-

b

"

"

"

"

"

"

"

"

"

"

tion Planning provides methods for effective decision making that are increasingly supported by automated tools and work-flows. It relies on descriptive information about the objects and carefully chosen samples [5] to conduct controlled experiments and analysis for the purpose of decision support and documentation. A key part of a preservation plan is the description of the collection [1]. This content profile does not describe the information held in the objects, nor their specific domain purpose (blueprints, newspaper articles, government emails, etc.). It serves a far more specific purpose: It aggregates and combines the meta data of the objects in order to give a better overview to the planner and help her understand the implications of the chosen preservation alternatives. Unfortunately, such meta data profiling is often neglected because of the lack of automation support and the scale of real digital object sets.

Consider a set of three pdf files where the meta data known is the identification data consisting of format and format version, as shown in Table 1. Most preservation experts would agree that file 1 and file 2 are similar. It is likely that the preservation risks of file 3 are handled differently. But consider the same three files with additional knowledge provided by deeper characterization, as shown in Table 2. Many experts will consider file 1 and file 3 to be homogeneous and may treat file 2 differently.

The problem is that in a real world scenario, such a collection will be significantly larger, with a complex format profile and many more characteristics. This makes it difficult to comprehend the differences of the objects and divide them into homogeneous sets based on those characteristics, that cause the issues during preservation actions. Only in-depth characterization can provide the necessary information required for effective planning. The goal of such a content profile thus is to provide comprehensive overview of a collection considered for long-term preservation.

2. STATE OF THE ART

Approaches and tools demonstrated thus far are generally restricted to format identification [2] or to small scales of content. Hitchcock et al. argue that although profiles can be based on many different aspects, the one that matters is the file format [4]. While it is one of the most significant properties within a collection, it often does not provide enough information to preserve it successfully. As in the example above, this most importantly applies to content that is homogeneous in terms of the format, where potential failures during the execution of a preservation action come from the subtleties of other characteristics.

Automatic meta data extraction is done by numerous tools,

Table 1: An example set of three pdf files

Characteristic	File 1	File2	File 3
Format	PDF 1.2	PDF 1.2	PDF 1.4

Table 2: The same set with additional meta data

Characteristic	File 1	File2	File 3
Format	PDF 1.2	PDF 1.2	PDF 1.4
Page count	20	20.000	40
Encryption	Yes	No	Yes
File Size	1 MB	120 MB	2 MB
Valid	No	Yes	No
Well-formed	Yes	Yes	Yes

such as Apache Tika, JHove and many more. The FITS¹ tool follows a different approach that unifies many different characterization tools but provides a normalized output of their results and gives indicators for their validity. These features provide a solid basis for preservation analysis and a complete content profile.

One key argument against the usage of in-depth characterization is that the analysis of meta data produced is extremely time-consuming. This stems from the observation that even the amount of meta data itself may be substantial. However, scalable approaches for content characterization can build on parallel architectures such as map-reduce to increase the processing speed in the analysis itself [3].

3. SCALABLE CONTENT PROFILING

Content profiling consists of three high-level steps: meta-data gathering, processing & aggregation and meta-data analysis. The first step transforms the data in a model that supports faster and scalable analysis and stores it. Post-processing solves issues, such as conflict resolution, due to the normalization of data provided by different tools and aggregation provides a machine readable overview of the data. The last part of profiling offers the planning expert a service on top of the data. It helps the analysis of the subtleties of the objects and partitioning the content into smaller sets fit for a specific preservation action.

Clever, Crafty, Content Profiling of Objects (c3po²) is a software tool prototype, which uses FITS generated data of a digital collection as input and generates a profile of the content set in an automatic fashion. It is designed in a way so that different meta data formats originating from other tools can be easily integrated. The tool follows the proposed three part profiling process and provides facilities for data export and further analysis of the content, such as helpful visualizations of the meta data characteristics, partitioning of the collection into homogeneous sets based on any known characteristic. In order to support the decision making it also makes use of different algorithms that choose a small set of sample records (up to 10) based on the size of objects, the distribution of specific characteristics, or other common features. For each chosen partition of the content, a special machine-readable profile can be generated that contains aggregations and distributions for many of the properties. The profile optionally contains the set of chosen representative samples as well as their identifiers within a content repository and a list of all objects that fall into the particular partition. A machine-readable content profile conforming to such a specific format plays an important role for integration with a planning component, content repositories and monitoring systems and thus for the automation of the entire cycle of planning and operations.

¹<http://code.google.com/p/fits/>

²<http://github.com/peshkira/c3po>

c3po makes use of a MongoDB³ document store, which is a scale-out NoSQL solution. This provides a huge performance boost in comparison to a traditional relational database solution, since the key-value structure of data closely corresponds to the underlying structure of the meta-data collected. The native map-reduce capabilities of this storage solution enable c3po to build format profiles, distributions of any other characteristic and combinations thereof in the order of seconds for hundreds of thousands objects.

To reduce network overhead, c3po offers a command line tool that can be executed near the data and the document store and also a web application, that can be deployed separately and used for visual aid to the planning experts.

4. RESULTS & OUTLOOK

Parsing, post-processing, aggregating, and generating a profile for a medium real world collection consisting of about 42 thousand FITS files (documents) currently takes about 1.5 minutes on a standard PC with 4GB RAM and 2.3 GHz CPU. Similar processing on a much larger real world content set from a web archive, consisting of about 1.3 million FITS files, completes in under 10 minutes on a single machine with 8 CPU cores. Filtering the content based on different characteristics is done via map-reduce jobs, and although it takes 30 to 60 seconds for each job, it turns out to be of great value during analysis.

Future research includes case studies of different content types and algorithms for more effective stratification of samples, integration with Plato⁴, as well as with repositories and automated monitoring services. Support and interfaces for different characterization tools and further tool optimizations will provide even more solid basis for faster and scalable analysis.

Acknowledgements

Part of this work was supported by the European Union in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

5. REFERENCES

- [1] C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *IJDL*, 10(4):133–157, 2009.
- [2] T. Brody, L. Carr, J. Hey, A. Brown, and S. Hitchcock. Pronom-roar: Adding format profiles to a repository registry to inform preservation services. *The International Journal of Digital Curation*, 2(2), November 2007.
- [3] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, Jan. 2008.
- [4] S. Hitchcock and D. Tarrant. Characterising and preserving digital repositories: File format profiles. *Ariadne*, (66), 2011.
- [5] H. Kulovits, A. Rauber, A. Kugler, M. Brantl, T. Beinert, and A. Schoger. From TIFF to JPEG2000? preservation planning at the bavarian state library using a collection of digitized 16th century printings. *D-Lib Magazine*, 15(11/12), 2009.

³<http://www.mongodb.org/>

⁴<http://ifs.tuwien.ac.at/dp/plato>

Defining Digital Curation through an Interactive, Informal Critical Delphi Approach

Lori Podolsky Nordland
McGill University
3459 rue McTavish
Montréal, QC, Canada
1 (514) 398-2955
Lori.Nordland@mcgill.ca

Carolyn Hank
McGill University
3661 rue Peel, Room 210
Montréal, QC, Canada
1 (514) 398-4684
Carolyn.Hank@mcgill.ca

ABSTRACT

Digital curation may be thought of as a set of strategies, technological approaches, and activities for establishing and developing trusted repositories, and ensuring long-term access to digital assets. It spans many disciplines and communities, as well as individuals seeking to maintain, preserve and add value to the ever-expanding body of digital content. This diversity has given way to ambiguity in defining digital curation, particularly in consideration of potentially synonymous terms, such as digital stewardship, preservation, and archiving. This poster will provide a forum for participants to challenge and engage in the dialogue that defines and describes digital curation.

Categories and Subject Descriptors

H.1.1 [Information Systems]: Systems and Information Theory – *information theory*

K.3.2 [Computers and Education]: Computer and Information Science Education – *curriculum*

General Terms

Management, Measurement, Documentation, Theory.

Keywords

Digital curation, preservation, archives, definition, consensus.

1. INTRODUCTION

Digital curation rose out of the limitations that were being found within digital preservation [1]. In a 2002 survey conducted on the meaning of preservation, respondents wrote that preservation had become antiquated when describing the actions and processes undertaken to preserve, manage and make accessible digital information, and subsequently suggested digital curation and digital stewardship as alternative terms [1]. In this context, multiple orientations exist as curation may be a process, a state of being for the record, or a statement about the properties of the record. These orientations transfer into the scope of the foundational framework along with the best practices and general procedures that make up the everyday duties of the curator.

Communities of digital curation professionals, including archivists, librarians, data scientists, and computer programmers, may perceive the key concepts underlying digital curation differently. These foundational building blocks provide the uniform conception within a discipline or study as it establishes the theoretical base [2]. In turn, these inform a consensus of operational terms that formulate a comprehensive definition spanning disciplines and communities. In the case of digital curation, the commonly cited definition comes from the Digital Curation Centre (DCC). The first part of the definition describes

what is digital curation; whereas, the latter paragraphs address the motivations and benefits of digital curation. The DCC website states the following:

Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle.

The active management of research data reduces threats to their long-term research value and mitigates the risk of digital obsolescence. Meanwhile, curated data in trusted digital repositories may be shared among the wider UK research community.

As well as reducing duplication of effort in research data creation, curation enhances the long-term value of existing data by making it available for further high quality research. [3]

The DCC was instrumental in focusing the direction and mandate of digital curation, providing, so to speak, the roots to this new field of research and practice. Playing upon the description of these roots, it may be helpful to visualize digital curation as a tree. The DCC's definition of digital curation [3] may be seen as the root, with synonymous terminology and alternative or derived definitions as the branches. Preceding and on-going research and practices form the concepts and principles of digital curation, which then moves towards, that is grows and matures, the theory on digital curation. This theory gives the tree its balance and shape, which references that important texture and context in understanding and describing digital curation.

2. RESEARCH NEED

The different branches of the digital curation tree have led to different conceptualizations of digital curation. In his research, Smith II demonstrated how various definitions led to ambiguity and confusion over the meaning of digital curation, and thus advocated for a “common nomenclature” [4]. He argued for stabilization in the meaning and scope of the term digital curation, which is important for efficient work and communication. A decade ago, Cloonan and Sanett also acknowledged a similar concern in their study of preservation strategies activities and the evolution of the definition of preservation [1]. One finding concerned the fluidity of the terminology, such as when digital curation and digital stewardship are used interchangeably [1]. A later study by Bastian, Cloonan and Harvey expanded on this interchangeability of terms in their examination of the etymology of digital curation and digital stewardship to best capture the scope of digital preservation, care and administrative responsibility of digital collections [5]. In both studies, the activities and functions have evolved beyond that of preservation,

and should account for resource management, access and the ability to present information, at the very least.

Lee and Tibbo [6] and Yakel [7] have provided alternate definitions of digital curation that do not explicitly include the concept of value-added. Lee and Tibbo describe digital curation in terms of the historical etymology of curation and the work and subsequent contribution of archivists [6]. In this example, the emphasis in the DCC definition on scientific data expands to that of cultural, social, historical and scientific material. Yakel omits both the term value-added and scientific data as she defines digital curation as the “active involvement of information professionals in the management, including the preservation, of digital data for future use” [7]. Beagrie [8] expands upon the DCC definition and incorporates the notion of adding value when he views the term as not only ‘being used for the actions needed to maintain digital research data and other digital materials over their entire lifecycle and over time for current and future generations of users,’ but also “all the processes needed for good data creation and management, and the capacity to add value to data to generate new sources of information and knowledge.” These examples highlight the trends that focus on key terms, such as “adding value” and “archiving” or “preservation” of digital assets. Yet concurrently, they also demonstrate how the operationalization of these terms is vague, potentially contributing to uncertainty in how to implement digital curatorial activities.

In respect to the variety of stakeholders working in the area of digital curation, representing different disciplines and communities, the foundational building blocks of the core definition for digital curation may be defined differently by practitioners, researchers and educators. Furthering the understanding of how key terms are used synonymously and in practice will aid in learning how the definition of digital curation is evolving. Additionally, while previous research has strongly focused on the root of digital curation, the branches of related professions has been limited, and thus opportunities for richer, contextual meaning and descriptions are still outstanding.

3. METHODOLOGY

In order to present the various definitions of digital curation, a formal literature review has been undertaken. The literature review reflects the perceptions of academic experts in the field of digital curation and information studies. Presented will be a brief summary on the emergence of the term, digital curation, and the key concepts and principles underlying it. This includes an examination of digital curation’s relationship to other related terms, such as digital preservation and digital archiving, and key concepts, such as value-added. In addition to the summary, the poster showcases other definitions of the term from various disciplines for iPRES attendees to review along with the core definition from the DCC.

The poster is intended for high interactivity. Through an informal Critical Delphi technique, to facilitate and moderate discussion,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Conference '10, Month 1–2, 2010, City, State, Country.
Copyright 2010 ACM 1-58113-000-0/00/0010 ...\$15.00.

iPRES attendees will be invited to share their own perspectives on the terminology around digital curation [2]. Participation will be voluntary, based upon those who visit the poster session. The community of the iPRES conference provides a diversified set of experts and professionals to further inform a diverse definition of digital curation. Their perspectives will be added directly to the poster, creating an interactive media on which to simultaneously gather, discuss, collaborate and analyze the concepts around digital curation, allowing participants to immediately comment on the emerging data and provide feedback during the poster session.

As the poster session progresses, key concepts and terms, including the foundation building blocks, will start to be categorized and indexed. These links will then be transferred to a concept map to highlight areas of commonality and divergence in the various definitions. The use of concept mapping had been used in similar projects. For example, the SHAMAN [9] project explored the literature and identified four categories of needs in digital preservation. These categories were then mapped to demonstrate the role information behaviour research played in the information system design. This level of participation provides greater opportunity for clarification of discussion during data collection, as well gaining feedback on both the data and the approach during the early stages of the research.

4. INTENDED OUTCOMES

A solid understanding of digital curation and an agreement on the foundational building blocks will lead to a cohesive definition. Consensus towards a cohesive definition will also be a strong tool in establishing clear objectives and promote a stronger identity and practices. Until then, the term digital curation is at risk of being misappropriated and, potentially, leading towards fragmentation within the professional and academic communities.

The end goal of this poster session will be to stimulate discussion and interest that moves towards a proposed, collective definition of digital curation. The data gathered during the poster session will be used to frame subsequent data collection. Following from the session, those attending will be asked if future contact will be permitted as the Critical Delphi technique employs a methodology in which information is gathered through a series of questionnaires, in which each subsequent round informs the next. This poster session will serve as a first, preliminary round of a planned, subsequent study to map an ontological tree of digital curation.

5. REFERENCES

- [1] Cloonan, M., and Sanett, S. 2002. Preservation strategies for electronic records: where are we now – obliquity and squint? In *American Archivist* 65 (2002), 70-106.
- [2] Zins, C. 2007. Conceptions of Information Science. In *Journal of The American Society for Information Science And Technology* 58 (3, 2007), 335–350.
- [3] Digital Curation Centre DOI= www.dcc.ac.uk.
- [4] Smith II, P. 2011. Defining digital curation understanding across disciplines, institutions, and organizations in the US. IASSIST 2011, 37th Annual Conference, Data Science Professionals: A Global Community of Sharing, DOI= http://www.iassistdata.org/downloads/2011/2011_poster_smith.pdf.

- [5] Bastion, J., Cloonan, M.V. and Harvey, R. 2011. From Teacher to Learner to User: Developing a Digital Stewardship Pedagogy. In *Library Trends*. (Winter), 607-622.
- [6] Lee, C. and Tibbo, H. 2011. Where's the Archivist in Digital Curation? Exploring the Possibilities through a Matrix of Knowledge and Skills. In *Archivaria*.72 (2011), 123-168.
- [7] Yakel, E. 2007. Digital curation. In *OCLC Systems & Services* 23 (4, 2007), 335 – 340.
- [8] Beagrie, N. 2006. Digital Curation for Science, Digital Libraries, and Individuals. In *International Journal of Digital Curation* 1 (1, 2006), 3-16.
- [9] SHAMAN DOI= <http://informationr.net/ir/15-4/paper445.html>

bwFLA – Practical Approach to Functional Access Strategies

Klaus Rechert, Dirk von Suchodoletz and Isgandar Valizada

Department of Computer Science
University of Freiburg, Germany

ABSTRACT

The goal of the bwFLA project is the implementation and development of services and technologies to address Baden-Württemberg state and higher education institutes' libraries' and archives' challenges in long-term digital object access. The project aims on enabling diverse user groups to prepare non-standard artifacts like digital art, scientific applications or GIS data for preservation. The project's main goal is to build-on ongoing digital preservation research in international and national projects to integrate workflows for emulation-based access strategies.

1. MOTIVATION

The Baden-Württemberg Functional Longterm Archiving and Access (bwFLA)¹ is a two-year state sponsored project transporting the results of ongoing digital preservation research into the practitioners communities. Primarily, bwFLA creates tools and workflows to ensure long-term access to digital cultural and scientific assets held by the state's university libraries and archives. The project consortium brings together partners across the state, involving people of university libraries and computer centers, library service facilities and archives providing a broad range of backgrounds and insights into the digital preservation landscape.

The project builds on existing digital preservation knowledge by using and extending existing preservation frameworks. It will define and provide a practical implementation of archival workflows for rendering digital objects (user access) in their original environment (i.e. application) with no suitable migration strategies available, like interactive software, scientific tool-chains and databases, as well as digital art. Thereby, the project focuses on supporting the user during object ingest to identify and describe all secondary objects required [?]. This way technical meta-data will be created describing a suitable rendering environment for a given digital object. The technical meta-data will serve as a base for *long-term access through emulation*.

¹bwFLA homepage, <http://bw-fla.uni-freiburg.de>.

2. PROJECT SCOPE

In most cases the best way to render a certain digital object is using its creating application, since those cover most of the objects' significant properties thus ensuring rendering of a better quality. Existence of alternatives is even not guaranteed in many cases due to the proprietary nature of the objects' file formats or its interactive nature. Preservation of the original environment is therefore crucial for the preservation of digital objects without suitable migration options, e.g. singular digital objects or digital art. The project develops workflows, tools and services required to safeguard future access of a digital object's rendering environment.

For current computer environments and applications plenty of user knowledge is available. More specifically owners of specific digital objects have good knowledge on the object's significant properties and their desired functions and utility. If such an object becomes subject to digital preservation, a defined workflows should support the preservation process of the object's rendering environment by

1. making use of the user's knowledge to identify all necessary components of the object's rendering environment such that the rendering environment is complete and there are no dependency conflicts,
2. preserving the knowledge on installation and configuration of the software components,
3. providing a preview of the emulated / recreated environment, such that the user is able to test if the chosen setup is meeting the desired rendering quality.

3. USE-CASE EXAMPLES

Archives and libraries keep digital objects like PhD theses since a few years and have new well established workflows e.g. to ingest PDFs into their collections and long-term storage. This procedure is often, at least partly, run by the contributor of the object. But what if the dissertation is complemented with an optical medium containing primary research data and the application which is able to render or interpret such data? Ensuring functional long-term access to such objects is a challenge. A similar problem is posed by digital art objects or GIS data, e.g. part of students' master theses or as an outcome of a research project.

4. WORKFLOWS

In order to describe a digital object's rendering environment technical meta-data has to be generated. This data

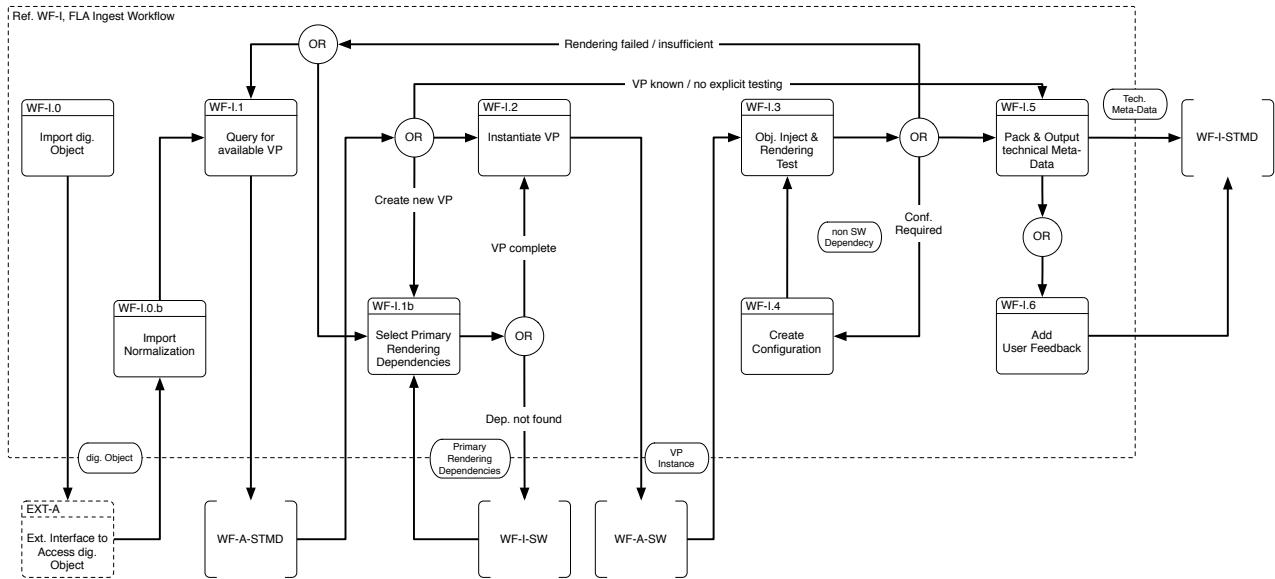


Figure 1: bwFLA ingest workflow; creating a description of a rendering environment for a given digital object.

will be generated through a constructive approach; the contributor is required to rebuild the objects original environment on a virtual or emulated machine. Through this guided process (e.g. ingest workflow) meta-data on the recreated environment is created in an automated way (cf. Fig. 1).

1. Relying on the contributor's knowledge of the object, the contributor chooses the *primary rendering dependencies*, which are known to render the digital object properly. If all or some dependencies can not be satisfied, the contributor is directed to the software-archive workflows to ingest missing software components.
2. In a second step the software environment containing the primary rendering dependencies is prepared either manually or in an automated way [2] and the digital object is prepared for transportation into the emulated environment.
3. Finally the user is able to access the digital object through an appropriate emulation component [1] to rate and approve the rendering quality. If the rendering result is signed off, the description of the rendering environment is available for the given object.

While this procedure involves a significant amount of manual labor, preservation planning costs will be reduced due to focusing on emulator software. Furthermore, by integrating a feedback loop with the original contributor, rendering quality and long-term access options may be guaranteed at ingest time.

5. CONCLUSION & OUTLOOK

After a number of successful national and international initiatives and projects on digital preservation and access it is time to leverage the results to an average memory institution having to deal with these matters. As the bwFLA

project is comparably small it focuses on the extension of existing workflows to enable efficient ways to open these processes to be compliant with more complex digital material delivered. Building on the basis of existing frameworks such as PLANETS and KEEP² encourages the project's sustainability.

In the project's first phase, a functional prototype for selected classes of digital objects will be delivered. Based on the experience gained, documentation and training material to enable a structured development of new workflows for future classes of digital objects will be provided.

Acknowledgments

The work presented in this publication is a part of the bwFLA project sponsored by the federal state of Baden-Württemberg, Germany.

6. REFERENCES

- [1] D. von Suchodoletz, K. Rechert, and I. Valizada. Remote emulation for migration services in a distributed preservation framework. In *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES2011)*, pages 158–166, 2011.
- [2] D. von Suchodoletz, K. Rechert, R. Welte, M. van den Dobbelen, B. Roberts, J. van der Hoeven, and J. Schroder. Automation of flexible migration workflows. *International Journal of Digital Curation*, 2(2), 2010.

²Currently maintained at <http://www.openplanetsfoundation.org> for follow up activities.

Will Formal Preservation Models Require Relative Identity?

An exploration of data identity statements

Simone Sacchi, Karen M. Wickett, Allen H. Renear
Center for Informatics Research in Science and Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel Street, MC-493
Champaign, IL 61820-6211 USA
{sacchi1,wickett2,renear}@illinois.edu

Keywords

Data, Identity, Scientific Equivalence, Data Curation, Digital preservation,

1. INTRODUCTION

The problem of identifying and re-identifying data put the notion of "same data" at the very heart of preservation, integration and interoperability, and many other fundamental data curation activities. However, it is also a profoundly challenging notion because the concept of *data* itself clearly lacks a precise and univocal definition. When science is conducted in small communicating groups, with homogeneous data these ambiguities seldom create problems and solutions can be negotiated in casual real-time conversations. However when the data is heterogeneous in encoding, content and management practices, these problems can produce costly inefficiencies and lost opportunities. We consider here the *relative identity* view which apparently provides the most natural interpretation of common identity statements about digitally-encoded data. We show how this view conflicts with the curatorial and management practice of "data" objects, in terms of their modeling, and common knowledge representation strategies.

In what follows we focus on a single class of identity statements about digitally-encoded data: "same data but in a different format". As a representative example of the use of this kind of statements consider the dataset "Federal Data Center Consolidation Initiative (FDCCI) Data Center Closings 2010-2013"¹, available at Data.gov. Anyone can "Download a copy of this dataset in a static format". The available formats include CSV, RDF, RSS, XLS, and XML. Each of this is presumably an encoding of the "same data". We explore three approaches to formalization into first order logic

¹<https://explore.data.gov/d/d5wm-4c37>

and for each we identify distinctive tradeoffs for preservation models. Our analysis further motivates the development of a system that will provide a comprehensive treatment of data concepts. [3].

2. PROBLEMATIC IDENTITY STATEMENTS

An example of the sort of statement we are considering is

$$\begin{aligned} a \text{ and } b \text{ are are the same data} \\ \text{but different XML documents} \end{aligned} \quad (\text{A})$$

Where "a" and "b" are identifiers or names of some sort and the object(s) they refer to are described as being different XML Documents but the same data, as would be for the RDF and XML files. The general form of such statements is:

$$x \text{ and } y \text{ are the same F but different Gs} \quad (\text{B})$$

Statements of this sort *relativize* identity (sameness) to particular categories such as, in this case, *data* or *XML Document* and imply that *x* and *y* are identical vis-a-vis one category (here, *data*), but different vis-a-vis another (here, *XML Document*). It is easy to see that the (B) may be understood as the conjunction of two clauses.

$$x \text{ is the same data as } y \quad (\text{C})$$

$$x \text{ is not the same XML Document as } y \quad (\text{D})$$

We now present three different approaches to understand these familiar sentence patterns.

2.1 The Classical View

The classical view asserts the principle known as Leibniz's Law (LL): if *x* and *y* are identical, then every property *x* has *y* also has. On the classical view this principle is a fundamental feature of our concept of identity and one that lies behind much ordinary reasoning; it is in fact an axiom in most formal logics that include identity. The classical view of identity will formalize (C) as follows:

$$\exists(x)\exists(y)(\text{data}(x) \ \& \ \text{data}(y) \ \& \ x = y) \quad (\text{1a})$$

This reads: "There exists an *x* and a *y* such that *x* is data and *y* is data and *x* is identical to *y*". On the Classical view *x* and *y* are the same "absolutely": if they are the same "data", they are the same (are identical) and so the

same with respect to any other possible characteristics. The classical view of identity will instead formalize (D) as follows:

$$\exists(x)\exists(y)(\text{XMLDocument}(x) \ \& \ \text{XMLDocument}(y) \ \& \ \neg(x = y)) \quad (1b)$$

This reads: “There exists an x and a y such that x is an XML Document and y is an XML Document and x is NOT identical to y . The function of the term “data” and “XML Document” is only to qualify the referents of x and y , not to describe the *kind* of identity asserted. Both (1a) and (1b) are ordinary expression in standard first order logic. On to this account, it follows from (1a) and (1b) that if x is data and y is an XML Document x is not the same thing as y . Yet there is “something” that *is* data and “something” that *is* an XML Document.

The classical view seems to imply that the natural analysis of our problematic identity sentences will result in a FRBR-like conceptual model with some number of closely related abstract entities — one of which is data, and another an XML Document — but no object that has all the properties that we seem to be implied in our ordinary colloquial sentences. This is the significance of our observing, above, that it is impossible for one thing to be both data and an XML Document, the conjunction of (C) and (D) is false for all values of x and y . Among the implications for data preservation is that if *data* is the actual target of preservation [3], we need to characterize it in terms that are independent, for example, of any specific file format. All approaches that rely on file-level definitions of data are fundamentally incomplete — if not flawed — and do not entirely support a correct representation of essential data transformations, like, for example, format migration.

2.2 Relative Identity View

Clearly the classical view does not respond to the sense of (A). The *relative identity* view was developed to accommodate the apparent semantics of these commonplace statements. According to the relative identity view x and y are identical only with respect to a general term (such as *data* or *XML Document*) that provides the *criterion* of identity [1]. Therefore a statement like “ x is identical with y ” is an incomplete expression, for which it “makes no sense to judge identity” unless we provide a criterion under which we can judge identity [1]. A consequence of this approach is that x and y can be *identical* with respect to some general count noun F, but *different* with respect to some other general count noun G. The relative identity view formalizes the conjunction of (C) and (D) like this:

$$\exists(x)\exists(y)((x =_{\text{data}} y) \ \& \ \neg(x =_{\text{file}} y)) \quad (2)$$

Although at first glance this view seems to match the grammar of how we often talk about digital objects, relative identity requires a new and very peculiar logical construct (an identity relationship that has three argument places: the terms identity is being applied to, and the sortal criterion). However, in a famous paper John Perry constructs a argument showing that relative identity is inconsistent with a number of very plausible assumptions², both at ontological

²See: <http://plato.stanford.edu/entries/identity-relative/>

and the logical levels [2]. From a modeling perspective, if we comply to *relative identity* we have also to abandon established paradigms such that of *levels of representation* that has proven to be a compelling modeling device to represent “what’s really going on” with preservation [3].

2.3 Equivalence Class View

A third view of identity statements such as (A) attempts to avoid the problems facing any analysis of identity by maintaining that, despite appearances, (A) is not really an identity statement at all, but rather an equivalence statement. According to the Equivalence Class View x and y may be different but *equivalent* with respect to specific equivalence relations. In our examples “data” and “XML Document” will both define equivalence relations: *data-equivalent* and *XMLDocument-equivalent* respectively. This view formalizes the conjunction of (C) and (D) like this:

$$\exists(x)\exists(y)((x \equiv_{\text{data}} y) \ \& \ \neg(x \equiv_{\text{XMLDocument}} y)) \quad (3)$$

We note that although (3) appears to use distinctive connectives it is plausible that they are best understood as *predicates*, therefore requiring no extensions to standard first order logic. The recently discussed notion of *scientific equivalence* [4] seems to reflect this approach. However, it leaves open the issue of a precise ontological representation of the entities involved in modeling digital objects for preservation.

3. CONCLUSION

We have drawn attention to a certain class of very important statements commonly made about scientific data in digital form. Although there are three plausible approaches to making logical sense out of these statements, the classical view of identity is decidedly superior to the others. The application of the classical view suggests the need for a system of distinct entities to correctly represent digitally-encoded data for preservation.

4. ACKNOWLEDGMENTS

The research is funded by the National Science Foundation as part of the Data Conservancy, a multi-institutional NSF funded project (OCI/ITR-DataNet 0830976).

5. REFERENCES

- [1] P. Geach. *Mental acts; their content and their objects*. 1957.
- [2] J. Perry. The same f. *The Philosophical Review*, 79(2):181–200, 1970.
- [3] S. Sacchi, K. Wickett, A. Renear, and D. Dubin. A framework for applying the concept of significant properties to datasets. *Proceedings of the American Society for Information Science and Technology*, 48(1):1–10, 2011.
- [4] C. Tilmes, Y. Yesha, and M. Haleem. Distinguishing provenance equivalence of earth science data. *Procedia Computer Science*, 4(0):548–557, 2011.

Training needs in digital preservation – A DigCurV Survey

Stefan Strathmann

Göttingen State and University Library
37070 Göttingen, Germany
++49 – (0)551/397806
strathmann@sub.uni-goettingen.de

Claudia Engelhardt

Göttingen State and University Library
37070 Göttingen, Germany
++49 – (0)551/3920284

claudia.engelhardt@sub.uni-goettingen.de

ABSTRACT

In this poster, we introduce the results of a survey of the training needs in digital preservation conducted by the DigCurV project.

Categories and Subject Descriptors

K.3.2 [Computers and Education]: Computer and Information Science Education – Accreditation, Computer science education, Curriculum, Information systems education

General Terms

Management, Human Factors.

Keywords

Digital preservation, digital curation, training, qualification, survey, questionnaire, needs assessment, vocational, cultural heritage institution.

1. INTRODUCTION

In 2011, the EU project Digital Curator Vocational Education Europe (DigCurV, <http://www.digcur-education.org>) conducted an online survey on training needs in digital preservation and curation. The study was part of the research DigCurV carried out to survey and analyze both the existing training opportunities and training needs in the cultural heritage sector. The results will be used to inform the development of a curriculum framework for vocational education and training in the field.

2. CONCEPTION

The online survey was carried out in July and August 2011. The target audience consists of staff members from libraries, archives, museums and other cultural heritage institutions as well as from organizations in the scientific and education sector, such as universities.

The questions addressed the digital preservation activities in which the respondents' organizations were engaged, the staff situation regarding this area, training plans for staff involved in the associated tasks as well as preferences in terms of methods and time frames for training, and whether training should be certified. Additionally, the respondents were asked to evaluate the

importance of a range of general as well as digital preservation-specific tasks and skills in terms of the work of staff involved in digital preservation and curation, and to assess the need for training with respect to several associated skills and competences.

3. SURVEY ANALYSIS

3.1 General information about the survey population

In total, 454 participants from 44 countries responded to the survey, the majority of them from Europe. The respondents were employed at a variety of institutions, mainly in the cultural heritage, the scientific and the education sectors, and were involved in digital preservation activities in manifold ways.

3.2 Involvement in digital preservation activities

A majority (approx. three quarters) of the respondents reported that their organizations are already storing digital materials for long-term preservation, and another 18% plan to do so in the future. However, the survey results show that many organizations lack staff to take care of the associated tasks. While some plan to hire new staff, the majority (57%) do not.

It can be assumed that in many cases the tasks associated with digital preservation will be assigned to existing staff, who will then need to acquire the necessary knowledge, skills, and competences. Hence, it seems very likely that there will be a considerable demand for corresponding training opportunities in the near future.

3.3 Training plans and preferences

Accordingly, many of the respondents' organisations are planning training for digital preservation staff. 35% of the respondents reported that there are plans to train staff with no previous experience in digital preservation, and 31% stated that there will be training for staff who already have some experience. Regarding the certification of training, the respondents' opinions were divided. Half responded that certification of training is important; the other half replied that such certification is not necessary.

In terms of the methods and time frames for training, the respondents indicated clear preferences. Small group workshops stood out as the method that was regarded as most suitable for their organisation by 75% of the respondents, followed by blended learning (a combination of face-to-face instruction and online components), which was chosen by 38%. The most popular time frame, one-time training events of 1-2 workdays, was chosen by 55% of the respondents. One-time events lasting 3-5 workdays

were the second most popular time frame, as indicated by about 30% of the participants.

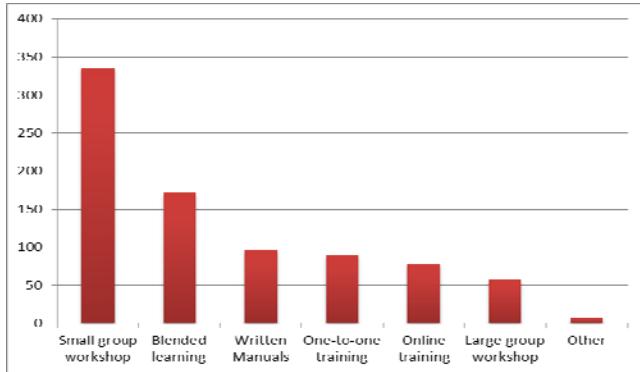


Figure 1. Most suitable training method¹

3.4 Skills and competences needed in digital preservation

With regard to the importance of several tasks and skills in terms of the work of staff involved in digital preservation and curation, the respondents indicated a high degree of relevance of both general skills and digital preservation-specific and technical skills. In terms of the latter, all of the listed items (Preservation Planning, Ensuring access, Managing data, Evaluating and selecting data for long-term preservation, Storing data, Ingesting data, Research, development and implementation of digital preservation environments, Administering the archive) were assessed to be either essential or important by more than 90% of the participants. As for general skills, three of them were regarded as being either essential or important by more than 95% of the survey population: collaborating with others, communicating with others, and an affinity for technology.

3.5 Training needs with regard to digital preservation and curation

The results of the assessment of the need for training with regard to the different skills and competences suggest a substantial need for both digital preservation-specific and technical skills and general skills. The percentage of respondents who reported either a great need or a moderate need were consistently very high for each of the given digital preservation-specific and technical skills: between 86% and 96%. The greatest need in this respect was expressed for general or basic knowledge of digital preservation issues, preservation and data management planning, and preservation tools. Regarding general skills, the numbers were lower but nevertheless still significant: between 60% and 85% of the respondents indicated a moderate or a great need for the several items given, with the greatest need stated for liaising between customer and information technology experts. When asked to prioritize the most pressing needs, they are clearly ascribed to the digital preservation-specific and technical skills (see Figure 2). The three items where the need was expressed to be most urgent were already mentioned above: general or basic

knowledge of digital preservation issues (chosen by 49% of the respondents), preservation and data management planning (48%) and preservation tools (38%).

4. OUTLOOK

The results of the survey show that digital preservation and curation is a field of activity that is becoming more and more relevant for cultural heritage as well as other institutions. However, many of these institutions are suffering from a lack of appropriately skilled staff to take care of the associated tasks. Arising from these circumstances is an urgent need for training that calls for immediate action. As a response to this situation, the DigCurV project is developing a curriculum framework for vocational education and training in the field. The design of the curriculum will be informed by the results of the training needs survey and of other research conducted by this project. The curriculum will be evaluated by trainers and also tested in practice. One opportunity to do so will be the nestor/DigCurV School event to be held in Germany in autumn 2012. In addition, DigCurV is actively seeking trainers and other projects to collaborate on the evaluation and the further development of the curriculum framework.

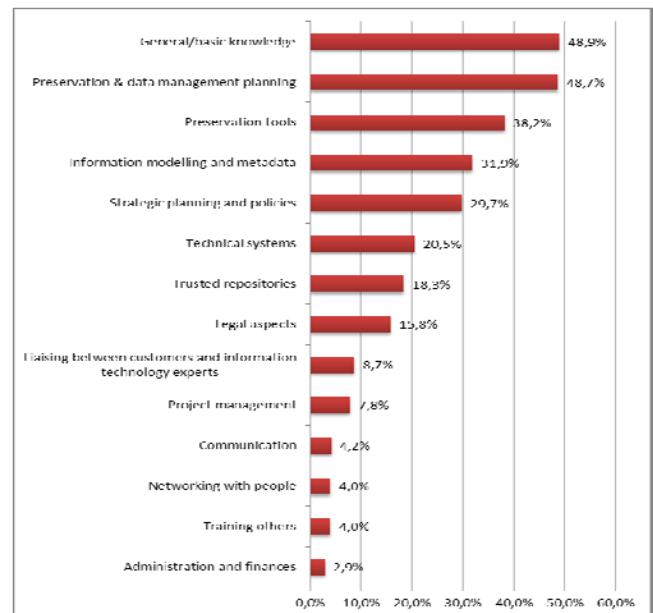


Figure 2. Most pressing needs for training²

5. REFERENCES

- [1] Karvelyte, V., Klingaite, N., Kupriene, J., Molloy, L., Snow, K., Gow, A. 2012. D2.1 Report on baseline survey and evaluation framework. Section 1: Training opportunities survey. <http://www.digcur-education.org/eng/content/download/3033/41768/file/D2.1.1%20DigCur-survey-into-training-opportunities.pdf>
- [2] Engelhardt, C., Strathmann, S., McCadden, K. 2012. Report and analysis of the survey of Training Needs.

¹ Figure 1 shows the results for the following question: "Which training methods do you consider the most suitable for your organisation?" 446 respondents answered this question; up to two answers were allowed.

² Figure 2 shows the results for the following question: "In which of the following digital preservation-related field/s is the need for training most pressing?" 448 respondents answered this question; up to three answers were allowed.

<http://www.digcur-education.org/eng/content/download/3322/45927/file/Report%20and%20analysis%20of%20the%20survey%20of%20Training%20Needs.pdf>

Retrocomputing as Preservation

Yuri Takhteyev
University of Toronto
140 St. George Street
Toronto, ON M5S 3G6
yuri.takhteyev@utoronto.ca

Quinn DuPont
University of Toronto
140 St. George Street
Toronto, ON M5S 3G6
quinn.dupont@utoronto.ca

ABSTRACT

This project explores the world of retrocomputing, a constellation of largely—though not exclusively—non-professional practices involving old computing technology. Retrocomputing includes many activities that can be seen as constituting “preservation,” and in particular digital preservation. At the same time, however, it is often transformative, producing assemblages that “remix” fragments from the past with newer elements or joining historic components that were never previously combined. While such “remix” may seem to undermine preservation, it allows for fragments of computing history to be reintegrated into a living, ongoing practice, contributing to preservation in a broader sense. The seemingly unorganized nature of retrocomputing assemblages also provides space for alternative “situated knowledges” and histories of computing, which can be quite sophisticated. Recognizing such alternative epistemologies in turn paves the way for alternative approaches to preservation. The institutional digital preservation community may have a lot to gain from paying closer attention to retrocomputing. This gain, however, should not just involve looking for ways to make use of the knowledge and labor of retrocomputing enthusiasts. Rather, it is important to recognize the value of their projects on their own terms and ask in what ways institutional efforts can support such projects.

Keywords

retrocomputing, software preservation, remix

In late March of 2012 Jordan Mechner received a shipment from his father, a box full of old floppies. Among them was a 3.5 inch disk labelled “Prince of Persia Source Code (Apple).” Mechner’s announcement of this find on his blog the next day took the world of nerds by storm. *Prince of Persia*, a game that Mechner developed in the late 1980s, revolutionized the world of computer games through its surprisingly realistic representation of human movement. After being ported to DOS and Apple’s Mac OS in the early 1990s the game sold 2 million copies.

Mechner’s original 1989 version, however, was written for Apple II, a platform already somewhat outdated at the time, and featured much more modest graphics and sound than the later DOS and Mac versions. This early version is still remembered—and played—by the aficionados, being easily available on the Internet in the form of disk image files derived from a “crack” of the game produced around 1990, credited to “The Crasher” and associates, and bearing a dedication to “Nebraska Cement Factory.”

The easiest way to run such images is to load them on one of the many Apple II emulators available online. For the more dedicated fans, however, there is the option of using original hardware. For some, this original hardware, is of course, Apple II. For others,

original hardware can mean *other* 1980s computers, including some that could not run the game at the time. For example, in 2011 a programmer known as “mrsid” successfully completed the project of porting the Apple II version of *Prince of Persia* to Commodore 64, a task that took him two and a half years. Projects such as mrsid’s would be much easier if the source code of the game were available. Yet, the code had long been presumed lost. Mechner’s discovery of the floppy thus generated much excitement.

The find, however, also presented a challenge. “I will now begin working with a digital-archeology-minded friend to attempt to figure out how to transfer 3.5” Apple ProDOS disks onto a MacBook Air and into some kind of 21st-century-readable format,” Mechner wrote on his blog. Mechner’s call for assistance brought two men to his door a few weeks later. One was Jason Scott, best known as the maintainer of textfiles.com, a website originally dedicated to preserving thousands of ASCII files shared on bulletin-board systems (BBS) in the 1980s and early 1990s, but then expanded to collect shareware CD images, audio files, and other digital artifacts from the era. The other man was Tony Diaz, a collector of Apple II hardware and the maintainer of the website apple2.org, dedicated to images of Apple II. Each man came with somewhat different tools. Scott brought DiscFerret, a small open-hardware device designed to read raw pattern of magnetism from a floppy, leaving the analysis and digitization of the pattern to a software tool, thus allowing flexible support for a wide range of approaches for storing data, as well as an ability to circumvent many antique copy-protection schemes. Diaz arrived with a van full of Apple II hardware—original, though rigged with substantial hardware and software modifications, including support for Ethernet, not available on the original Apple II.

With their help, Mechner’s original files were transferred to his MacBook Air, in a live-tweeted session tagged “#popsource” that attracted so much attention that Mechner’s website collapsed from the traffic. The source code was then quickly made available on GitHub, a site widely used for sharing open source code. Within hours, GitHub user “st3fan” made a modification commenting out the copy-protection code. This move was purely symbolic, since the posted code was incomplete at the time and could not actually be compiled and run. A few days later, however, a programmer working on an Apple II emulator credited the posted source code as a source of information that helped improve the emulator.

The story presented above provides a glimpse into the world of retrocomputing, a set of diverse practices involving contemporary engagement with old computer systems, which we have explored through a year long study combining online observation of retrocomputing projects and *in situ* interaction with the participants. Such practices are primarily private and non-professional, though this is not always the case—there is also a substantial economy providing products and services. And to the extent that retrocomputing participants are “hobbyists,” in the

sense of not paid for their work, they are hardly unskilled amateurs. Rather, their practice often demonstrates deep sophistication. In other words, many of them are “hobbyists” only in the same sense as many of the contributors to open source software, which today underlies much of the world’s computing infrastructure. Retrocomputing often involves old games, yet many participants also engage with non-gaming technology.

Many of the activities that make up retrocomputing can be seen as constituting collection and preservation, and many retrocomputing enthusiasts in fact recognize preservation of computer history as one of their key goals. Such activities involve efforts to collect and restore old hardware, develop emulators, and build substantial collections of old software. For example, it does not take long to find on the Internet disk images for *Prince of Persia* for Apple II, as well as a variety of emulators that can run them. Some of the emulators are produced by open source projects in the full sense of the term, many are “not quite open source,” for example distributed under licenses that prohibits commercial use. The differences highlight the distinct historic origins of retrocomputing and free software and the need to recognize retrocomputing itself as a historically situated practice.

Emulation requires images of original software. This means images of the original system software, as well as the application software that is to be emulated. Images collected and circulated by the hobbyists come from sundry sources. Some are actually quite old: the most commonly available image for the Apple II *Prince of Persia* appears to have originated around 1990. Some are more recent, but still produced by running image ripping software on old machines—or at least older machines. For example, one can read Apple II disks using early PC hardware with special software. This method can be quite challenging due to copy protection, as well as the gradual disappearance of older hardware. Perhaps the most sophisticated solution for this problem is exemplified by DiscFerret and KryoFlux—both hardware solutions that sit between a floppy disk drive and a contemporary computer, allowing the latter to scan the raw pattern of magnetization from a disk’s surface, defeating many of the copy-protection methods employed in the 1980s. Both projects are run by private groups, in case of DiskFerret—as an open source and “open hardware” project.

At the same time, closer attention to those retrocomputing projects reveals that they cannot be easily understood as just a matter of preservation in the narrow sense of keeping objects from the past fixed in their “original” form. Instead, retrocomputing is often transformative and involves construction of assemblages that “remix” fragments of old systems with newer elements, such as old software running on freshly debugged emulators or original hardware enhanced with contemporary networking. It can also involve a mixture of historic components that were never combined in the past, as in the case of mrsid’s porting of *Prince of Persia* to Commodore 64.

We conceptualize these transformative aspects of retrocomputing as a form of “remix”—a term popularized by Lessig (2008). Like the closely related concept of “collage,” the term “remix” refers to a creative and often playful reassembly of fragments of earlier works into something new. While the reasons for chimeric assemblages described above are sometimes pragmatic, at other times they are simply playful, carried out for fun. At a gathering of Commodore enthusiasts attended by one of the authors, a participant demonstrated an old Commodore 64C system that he had skillfully painted bright blue. He explained that the computer

was originally made of white plastic, but had turned an “ugly” yellow over time. Repainted blue, it looked “awesome.” Quite often, though, the pursuit of fun and beauty cannot be easily separated from the “pragmatic” motivation for remixing fragments of old computing. Much like Linus Torvalds describing his development of Linux as “just for fun” (Torvalds 2001), this notion of fun usually means getting satisfaction in finding solutions to technical problems, thus fusing “pragmatic” and “playful” motivations.

Playful remix inherent in much of retrocomputing may at first blush seem to be in contradiction to efforts preserving the history of computing. This contradiction, however, dissipates with further analysis. Even in the seemingly extreme case of re-painting an old machine to a new color—a step that cannot be undone—the work is preservative in that it restores the “awesomeness” that the artifact once possessed. A machine that was once a source of joy becomes capable of bringing joy once again. More generally, the remix inherent in retrocomputing allows continuous reintroduction of elements of past computing systems to ongoing, living practice. We understand practice as the system of activities comprised of people, ideas, and material objects, tied by shared meanings and joint projects (see also Takhteyev 2012). Computing artifacts are born into such systems and have power and meaning because of their linkages to other elements. Over time, however, some elements of these systems disintegrate or enter into new relationships and abandon the old ones. With the dissolution of relationships comes the fading of tacit knowledge that once made its use possible. Such processes of social decomposition may often be much more damaging to old computing systems than the physical breakdown of the hardware or storage media, and it cannot be stopped by isolating the fragments and shielding them from sunlight or improper temperature and humidity.

The decay of the socio-technical practice in which computing is embedded is partly stopped or even undone in retrocomputing, as ancient fragments are reintegrated into ongoing activities, becoming part of a *contemporary* living practice. Such integration allows for maintenance and even recovery of tacit knowledge (see also Galloway 2011), as well as continuous circulation of tools and resources. Retrocomputing is undergirded by a complex ecology of commercial, hobby, and grey market products and services, and it is this complex, interrelated ecosystem that allowed Mechner’s call to be answered so quickly, and his files to be transferred with such seeming ease from 1980s floppies to GitHub, where they will be both preserved and further remixed.

However, appreciating retrocomputing just for the resources it can provide may miss its deeper value. The practice of retrocomputing also provides space for ongoing circulation of meaning and divergent “situated knowledges” (Haraway 1988) and for alternative histories of computing. Consequently, it may not only provide us with new insights into the *how* of digital preservation, but also into the *what* and *why*. We may therefore need to recognize the value of retrocomputing projects on their own terms and look for ways to provide them with support.

ACKNOWLEDGMENTS

This paper is based on work supported by the Social Science and Humanities Research Council of Canada.

REFERENCES

- Galloway, P. (2011). Retrocomputing, archival research, and digital heritage preservation: a computer museum and iSchool collaboration. *Library Trends*, 59(4), 623–636. doi:10.1353/lib.2011.0014
- Haraway, D. (1988). Situated knowledges: the science question in feminism and the privilege of partial perspective. *Feminist Studies*, 14(3), 575–599. doi:10.2307/3178066
- Lessig, L. (2008). *Remix*. London: Bloomsbury Publishing PLC. Retrieved from http://www.bloomsburyacademic.com/view/Remix_9781849662505/book-ba-9781849662505.xml
- Takhteyev, Y. (2012) *Coding Places: Software Practice in a South American City*. MIT Press.

DuraCloud, Chronopolis and SDSC Cloud Integration

Andrew Woods

DuraSpace

165 Washington Street, Suite #201
Winchester, MA 01890
011-1-781-369-5880
awoods@duraspace.org

Bill Branan

DuraSpace

165 Washington Street, Suite #201
Winchester, MA 01890
011-1-781-369-5880
bbranan@duraspace.org

David Minor

UC San Diego

San Diego Supercomputer Center
La Jolla, CA 92093
011-1-858-534-5104
minor@sdsc.edu

Don Sutton

UC San Diego

San Diego Supercomputer Center
La Jolla, CA 92093
011-1-858-534-5085
suttond@sdsc.edu

Michael Burek

National Center for Atmospheric

Research

P.O. Box 3000
Boulder, CO 80307
011-1-303-497-1202
mburek@ucar.edu

ABSTRACT

In this paper we describe the interaction of three different systems: DuraCloud, a cloud-based service provider, Chronopolis, a digital preservation network, and the San Diego Computer Center's cloud service. This interaction is targeted to developing a new storage and preservation service available to a wide range of users.

General Terms

Algorithms, Management, Design, Experimentation

Keywords

Digital preservation, cloud storage, integration

1. INTRODUCTION

Since late 2011, Chronopolis¹, the San Diego Supercomputer Center (SDSC) Cloud Storage², and DuraCloud³ have been collaborating in an effort to add another layer of reliability to the field of distributed digital preservation. Each of our services have been designed to address a set of needs within the preservation community. Together, we have developed a single service that combines the archiving sensibilities of Chronopolis, the cost-effective, academic cloud storage of SDSC, and the provider-neutral access and preservation capabilities of DuraCloud. This paper will describe the details of the integration as well as follow-on activities. We will start, however, with a brief introduction to each of the constituent pieces.

2. INTEGRATION OVERVIEW

The DuraCloud/SDSC/Chronopolis integration was conceived of as a way to bridge the cost-effective dark archive, Chronopolis, with the online, dynamically accessible, provider-independent, preservation platform of DuraCloud. Prior to this effort, DuraCloud provided a mediation layer over three underlying commercial cloud storage providers: Amazon S3, Rackspace CloudFiles, and Microsoft Azure. The goals of the integration were to (1) add an academic cloud store (SDSC Cloud Service) to this list of providers supported by DuraCloud as well as to (2) enable DuraCloud users to replicate content to a geographically distributed, TRAC certified, preservation network (Chronopolis). Among other benefits, this integration supports the preservation strategy of distributing content across multiple geographic, platform, and administrative domains.

The first step in the integration process was to ensure that DuraCloud had the ability to store and retrieve content from the

¹ <http://chronopolis.sdsc.edu>

² <http://cloud.sdsc.edu>

³ <http://duracloud.org>

SDSC Cloud Service. This initial integration required very little effort due to the fact that the SDSC Cloud exposes what is emerging as the standard cloud storage interface, OpenStack's Object Storage API. Since this is the same API offered by an existing storage provider supported by DuraCloud, Rackspace Cloudfiles, the connector code was already in place. As a result, adding SDSC as an additional underlying storage provider to DuraCloud was as simple as providing a new connection URL.

While the integration between DuraCloud and SDSC Cloud was simple, the connection to Chronopolis required more care. The model of programmatic interaction with Chronopolis is very different from that of the common cloud storage providers, and as such a means of communication between the two systems needed to be defined. The final approach defines an asynchronous RESTful integration. Over the course of several months, a joint team with representation from all three organizations (SDSC, Chronopolis, and DuraSpace) created the set of calls required in the REST API. This work defined a series of steps which would be used to move content from the SDSC Cloud to Chronopolis and back as needed, all managed by calls made from DuraCloud to the REST API.

To move content from DuraCloud to Chronopolis, DuraCloud stores content in one or more SDSC cloud storage containers then sends a request to Chronopolis to read content from those container(s). Part of this request is a manifest file detailing each content item to be transferred. Chronopolis then pulls the requested content into its data centers and compares the file checksums with the provided manifest to ensure that all content was pulled successfully. Once the transfer is validated the objects are arranged in BagIt format⁴ and ingested into the Chronopolis system. The SDSC cloud service also allows custom meta name-value parameters to be assigned to objects. Using the manifest file, Chronopolis queries the SDSC cloud for any custom meta parameters and stores them with the ability to restore them if a retrieval is requested.

To retrieve content from Chronopolis, DuraCloud requests the restoration of all (or a subset) of the content back to an SDSC container, and Chronopolis performs the work of transferring the content from its data centers back to the SDSC Cloud. The inter-system communication is achieved via a REST server hosted by Chronopolis that receives requests from DuraCloud. (It should be noted that the Chronopolis REST server does not need to know that the client is a DuraCloud process. In this way, it is expected that other external systems could integrate with Chronopolis using the same methods.) The Chronopolis process behind the REST server is granted read/write access to one or more SDSC Cloud storage containers that are owned by DuraCloud.

The following three scenarios are covered by this integration: (1) A DuraCloud user wishes to have a snapshot of their content replicated to the Chronopolis preservation network. (2) A DuraCloud user wishes to restore a previously snapshotted collection from Chronopolis back to DuraCloud. (3) A DuraCloud user wishes to restore a single item of a previously snapshotted collection from Chronopolis back to DuraCloud.

3. NEXT STEPS

Due to the initial success of the DuraCloud/SDSC/Chronopolis integration a series of follow-on tasks are in process. Several end-to-end tests have proven the communication and data flow patterns. The objectives of the second round activities are to tease out any performance or technical issues as well as to discover and add any usability features that will ultimately ready the integrated system for production use.

On the technical side the next tasks will address security, inter-process communication, and performance improvements. The team will be layering security over the REST server in the form of SSL coupled with Basic-Auth. Beyond security, the API will also be extended to support a more robust back-flow communication mechanism. For example, after a content restore from Chronopolis to DuraCloud, if DuraCloud detects an error in the resultant file(s) an API method should be available to communicate that error back to the Chronopolis system. From a performance perspective we will be stressing the system to ensure that response times do not suffer at scale. We are in the process of staging a series of tests to back up and restore half a million data files up to one gigabyte in size.

As a step towards validating the capability and usability design of the integration, a set of interested institutions using DuraCloud will be invited to participate in beta testing. From the beta testing phase we expect to uncover any use cases that were not revealed in the earlier testing. Additionally, we hope to gain feedback on the general process flow and user interface. Assuming a successful period of beta testing, the expectation is that the SDSC and Chronopolis services nested under DuraCloud will be made publicly available as a production offering in the Fall of 2012.

In summary, the recognition that cloud-based, distributed, digital preservation is an increasingly emerging need, the three related technologies of Chronopolis, SDSC Cloud Service, and DuraCloud have undertaken the joint effort to provide the preservation and archiving communities with an additional layer of assurance. Not only will users be able to now select an academic cloud store in addition to existing commercial stores, they will also have the option to mirror their holdings through a dark archive network spanning across North America.

⁴ <https://wiki.ucop.edu/display/Curation/BagIt>

Demo – An Integrated System-Preservation Workflow

Isgandar Valizada, Klaus Rechert, Dirk von Suchodoletz, Sebastian Schmelzer
Chair in Communication Systems
University of Freiburg, Germany

ABSTRACT

Preserving a full computer system for long-term access is in some cases the option of last resort. In these situations the system consists of complex workflows and tool-chains paired with custom configuration such that the dismantling of individual components is too time consuming or even impossible to be carried out. For such situations a defined *system-preservation workflow* is required, ideally integrated into a digital preservation framework to ensure future rendering ability and quality.

This demonstration presents an integrated system-preservation workflow, designed to be performed by non-technical users and to ensure long-term access (e.g. through emulation) with guaranteed system properties. Furthermore, the workflow leverages a distributed framework, enabling on-site preservation tasks, while making use of a common shared knowledge base and ready-made access preview technologies.

1. MOTIVATION

In certain situations system-preservation of a whole computer system is inevitable. For instance in a scientific environment certain computer workstations have grown over time. During a perennial research a project's fluctuation of personnel is common and therewith system knowledge is volatile. Complex software workflows have been created, usually involving tailored software components and highly specialized tool-chains paired with non-standard system tweaks. Rebuilding the system from scratch is a time-consuming task involving manual labor, and it requires a technically skilled operator. Thus, preserving the complete workstation is more economical and, if carried out properly, full functionality of the system can be retained with minimal effort.

Through a system-preservation process an image of the complete content of a computer's hard disk is made. This image, a virtual hard disk, can then be run again with virtual hardware, i.e. virtualization or emulation technology. While the technical problems of imaging as well as re-enacting of

the workstation are solved in principle [1], a practical and usable solution including defined workflows and framework integration are still missing. Emulation projects like KEEP primarily focused on single object workflows and not yet providing the required functionality for system preservation.

The challenges of system-preservation tasks are manifold and mostly of technical nature. The tasks to be carried out require technical expertise on the targeted system, e.g. booting the system in a read-only mode to prevent system modifications and inconsistencies during the imaging process. Furthermore, the image needs to be post-processed and enriched with meta-data describing its original hardware environment.

To re-enact and test the preserved system image, knowledge of current emulation technologies is necessary. This knowledge may also include pre-processing steps to be carried out on the original system before imaging. Such tasks may include demotion of the system to default graphics drivers or disabling of external dependencies during the boot-process (e.g. mounting network shares, connections to license servers, etc.). Such external dependencies may be restored in the emulated environment in a post-processing step.

Therefore, the goal of providing a framework with an integrated workflow for system preservation is to enable non-technical users to prepare and perform a system preservation task and finally test the result produced before submitting the data to a digital long-term preservation repository. Thus, provided a successful run of the *system-preservation workflow*, a functional image of a real-life computer system is guaranteed to run on current emulator software. For long-term accessibility proper preservation planning is only required on emulation software. By providing immediate feedback, the owner of the preserved computer system is able to verify whether the subjective and objective significant properties were preserved.

2. TECHNICAL WORKFLOW

The system-preservation workflow and associated tools are an integrated part of the bwFLA framework,¹ designed to support the user in functional preservation tasks, especially leveraging emulation for object access. This demonstration describes the workflows and tools for system-preservation. As the framework does not provide a dedicated repository for digital object storage, the framework is de-

¹Developed as a part of the ongoing Baden-Württemberg Functional Longterm Archiving and Access project, <http://bw-fla.uni-freiburg.de>.

signed to interface with OAIS compliant DP frameworks within ingest and access workflows.

The workflow is split into three stages: First, the workstation subject to preservation is prepared and tailored software for the imaging process is generated. In a second step the imaging process is carried out and finally the generated image is tested and enhanced with meta-data.

Preparation

1. In a first step the user characterizes the computer system to be preserved (in the following denoted as "target system") by choosing one of the yet supported operating systems and computer architecture. Handling of not yet supported operating systems will be discussed below.
2. To perform the imaging process, the target system requires certain technical functionality's, e.g. USB-port or optical reader (CD-/DVD) and the ability to boot removable media. Furthermore, a (standard) network adapter is required. Following, the user is interactively questioned if the target system meets these requirements. Depending on the choices made, the imaging process is prepared either to be carried out on the target system, or on a different (modern) computer system. The latter option requires dismounting the hard-drive of the target system.
3. To ensure better compatibility with hardware emulator software a knowledge-base on different operating systems regarding their compatibility with emulators and hardware dependencies is part of the bwFLA system-preservation framework.
The user is presented with known issues based on his previous selections and step-by-step guides describing user-actions to be carried out on the target system. Common examples may include reducing screen resolution, reducing hardware driver dependencies or preparing network connections.
4. Based on the user's choices a specially tailored bootable image is generated and made available for download. The bootable image is either to be written on a USB pen-drive or CD/DVD. The boot-media contains a tailored Linux kernel suitable for booting on the target device, network configuration, and necessary credentials to connect to the bwFLA framework, e.g. to stream/upload the generated image in an automated way.

System Imaging

At this step the client uses the newly created bootable medium to boot the machine on which the preservation workflow is to be performed. After booting the bwFLA imaging system on the target system an automated process starts the imaging process, the gathering of the relevant hardware information about the target machine, and the uploading of this data to the bwFLA framework.

The only interactive choice allows the user to select the drive to be preserved, if multiple options are available. By default the standard boot-device is chosen. Currently, only complete block device preservation is supported. However, the ability of selective partition preservation is planned for



Figure 1: Target system booted from USB pen drive starting the system preservation process.

future work. During the imaging process the user is able to follow the progress on the Web-page that is monitoring the upload of the image data.

Verification and Submission

In a final step the disk image generated is post-processed (if required) for an appropriate emulator type, chosen based on selection made during the preparation process and information gathered during the imaging process.

Finally an bwFLA emulation component is invoked with the preserved system image and presented to the user. If the user approves the result, the image is submitted together with generated technical meta-data to a DP repository for further processing.

3. RESULTS & OUTLOOK

The workflow has been tested mainly on x86-based systems, mostly due to availability of high-quality emulators. However, adding new, yet unsupported systems is easily possible as long as the original workstation is able to boot a Linux kernel and connect to the network. In cases of older (potentially highly integrated systems) such as Notebooks without CD-ROM drives, USB-boot capabilities and even without network connection the external imaging workflow remains as a working option. However, in such a case the description of the original system remains to the user and has to be carried out manually.

With the integration of system-preservation workflows into a distributed digital preservation framework, a common knowledge base on preparing, imaging and re-enacting ancient computer system can be built, thus providing step-by-step instruction even to non-technical users. Having the feedback loop in available, the owner of a workstation subject to system-preservation is able to describe and approve desired properties of the system for later access.

4. REFERENCES

- [1] D. von Suchodoletz and E. Cochrane. Replicating installed application and information environments onto emulated or virtualized hardware. In *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES2011)*, pages 148–157, 2011.