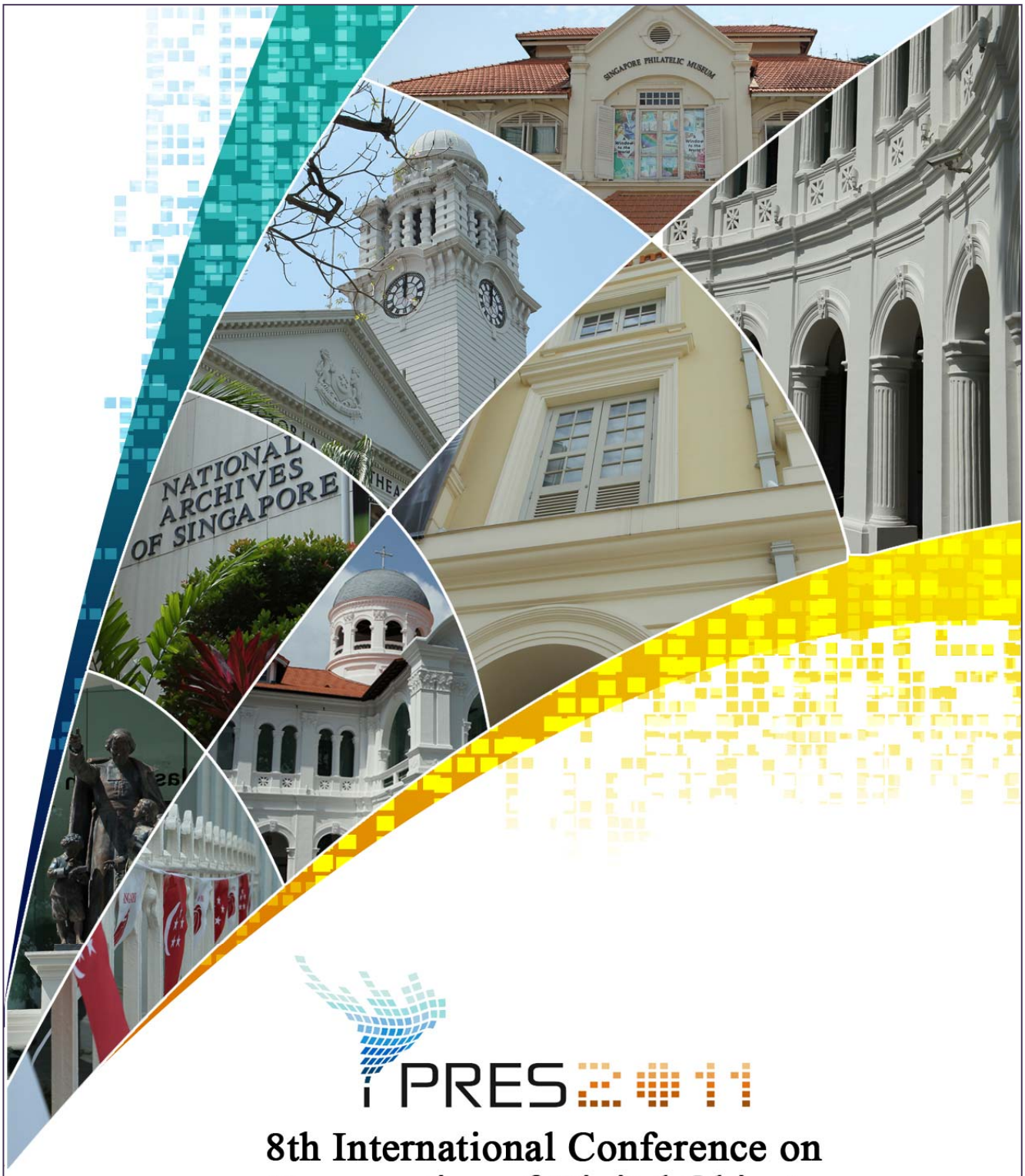


iPRES2011

8th International Conference on Preservation of Digital Objects

NLB | National Library Singapore

NANYANG TECHNOLOGICAL UNIVERSITY Wee Kim Wee School of Communication & Information

Blank page

8th International Conference on Preservation of Digital Objects

Editors: José Borbinha, Adam Jatowt, Schubert Foo,
Shigeo Sugimoto, Christopher Khoo, Raju Buddharaju

iPRES 2011 – 8th International Conference on Preservation of Digital Objects

National Library Board
100 Victoria Street
Singapore 188064

Nanyang Technological University
Wee Kim Wee School of Communication & Information
31 Nanyang Link
Singapore 637718

Cover design: Wu Zumei

ISBN 978-981-07-0441-4

PREFACE

iPRES2011, jointly organized by the National Library Board (NLB) Singapore and the Wee Kim Wee School of Communication & Information, Nanyang Technological University, Singapore is eighth in the series of the International Conference on Preservation of Digital Objects. The inaugural conference was held in Beijing, China in 2004. It has since travelled to Göttingen, Germany (2005), Ithaca, New York, U.S.A. (2006), Beijing, China (2007), London, U.K. (2008), San Francisco, U.S.A. (2009) and Vienna, Austria (2010). Originally to be held in Tsukuba, Japan in 2011, a decision was made to relocate the conference to Singapore following the massive earthquake and tsunami that struck the northeastern part of Japan on 11 March 2011, which caused serious damages and uncertainties in several infrastructural issues that arose thereafter.

Despite this change, a reduced working time frame and a formation of a new Organising Committee, we are pleased to note that the conference has continued to garner the strong interest and support of the preservation community. From a total of 72 submissions, 20 full papers, 13 short papers, 15 poster papers and 4 demonstration papers were selected by the Program Committee for presentation at the conference across 13 technical sessions on Governance, Migration, Emulation, Infrastructures, Risk and Planning, Cases and Communities, Cost Models and Complex Scenarios, and a poster-demonstration session.

The conference featured three panels to discuss important and timely preservation issues that included "What is worth saving? Selection and curation in web archiving" chaired by Lori Donavan, Internet Archives; "Preserving Web Archives" chaired by Raju Buddharaju, National Library Board Singapore and "Digital preservation and disaster scenarios" chaired by José Borbinha, IST/INESC-ID, Portugal.

The conference keynotes were delivered by Seamus Ross from the Faculty of Information, University of Toronto, Canada on the topic of "Digital preservation: Why should today's society pay for the benefit of society in future?", Ross Wilkinson from the Australian National Data Service (ANDS) on the topic of "Opportunities and challenges of preserving research data", and Mick Newnham from the National Film and Sound Archive, Australia on the topic of "Preserving motion picture film, so much to do so little time ....".

The preceding tutorials "Preservation metadata in PREMIS" by Raju Baddharaju & Haliza Jailani (National Library Board, Singapore) and Peter McKinney (National Library of New Zealand), and "Archiving websites" by Paul Wu (SIM University, Singapore) provided an opportunity to address a number of important topics in preservation. The conference hosted two timely workshops on "Steps toward international alignment in digital preservation" organized by Cal Lee, University of North Carolina, Chapel Hill, and "Web analytics" organized by the International Internet Preservation Consortium (IIPC).

In addition to the technical sessions, conference participants had the opportunity to further network at the conference dinner, join in a heritage trail and engage in site visits to the National Archives of Singapore and the Heritage Conservation Centre.

Finally, we wish to thank all conference participants, sponsors, exhibitors, the Local Committee and co-operating agencies including the Ministry of Information, Communication and the Arts, National Heritage Board and the Singapore Tourism Board, all of whom have contributed to the success of this conference.

Schubert Foo
Shigeo Sugimoto
José Borbinha
Adam Jatowt
Christopher Khoo
Raju Buddharaju

CONFERENCE ORGANIZATION

General Co-Chairs
Schubert Foo (Nanyang Technological University, SG)
Shigeo Sugimoto (University of Tsukuba, JP)

Programme Committee Co-Chairs
José Borbinha (IST / INESC-ID, PT)
Adam Jatowt (Kyoto University, JP)

Tutorial Chair
Masaki Shibata (National Diet Library, JP)

Workshop Chair
Natalie Pang (Nanyang Technological University, SG)

Publicity Co-Chairs
Joy Davidson (Digital Curation Centre, UK)
Christoph Becker (Vienna University of Technology, AT)

Local Organization Co-Chairs
Raju Buddharaju (National Library Board, SG)
Christopher Khoo (Nanyang Technological University, SG)

Programme Committee Members
Reinhard Altenhoener (German National Library, DE)
Bjarne Andersen (Statsbiblioteket, DK)
Andreas Aschenbrenner (State and University Library Goettingen, DE/US)
Thomas Baker (Dublin Core Metadata Initiative, DE/US)
Christoph Becker (Vienna University of Technology, AT)
Karim Boughida (George Washington University, US)
Adrian Brown (National Library of Australia, AU)
Gerhard Budin (University of Vienna, AT)
Priscilla Caplan (Florida Center for Library Automation, US)
Gerard Clifton (National Library of Australia, AU)
Euan Cochrane (Archives New Zealand, NZ)
Panos Constantopoulos (Athens University of Economics and Business, GR)
Paul Conway (University of Michigan, US)
Angela Dappert (British Library, UK)
Joy Davidson (University of Glasgow, UK)
Michael Day (UKOLN, University of Bath, UK)
Janet Delve (University of Portsmouth, UK)
Raymond Van Diessen (IBM, NL)
Jon Dunn (Indiana University, US)
Miguel Ferreira (University of Minho, PT)
Ellen Geisriegler (Austrian National Library, AT)
David Giaretta (Rutherford Appleton Laboratories, UK)
Andrea Goethals (Harvard University, US)
Emily Gore (Clemson University, US)
Mariella Guercio (Universita' degli Studi di Urbino Carlo Bo, IT)
Mark Guttenbrunner (Vienna University of Technology, AT)
Jane Hunter (University of Queensland, AU)
Angela Di Iorio (Fondazione Rinascimento Digitale, IT)
Greg Janée (University of California at Santa Barbara, US)
Leslie Johnston (Library of Congress, US)

Max Kaiser (Austrian National Library, AT)
William Kehoe (Cornell University, US)
Ross King (Austrian Institute of Technology, AT)
Amy Kirchhoff (Portico, US)
Hannes Kulovits (Vienna University of Technology, AT)
Brian Lavoie (OCLC, US)
Christopher A. Lee (University of North Carolina, US)
Bill Lefurgy (Library of Congress, US)
Jens Ludwig (Göttingen State and University Library, DE)
Maurizio Lunghi (Fondazione Rinascimento Digitale, IT)
Julien Masanes (European Web Archive, NL)
Nancy McGovern (ICPSR, US)
Andrew McHugh (HATII at University of Glasgow, UK)
Carlo Meghini (CNR- ISTI, IT)
Ethan Miller (University of California at Santa Cruz, US)
David Minor (University of California at San Diego, US)
Reagan Moore (University of Chapel Hill, NC, US)
Jacob Nadal (University of California at Los Angeles, US)
Heike Neuroth (Göttingen State and University Library, DE)
Quyen Nguyen (National Archives and Records Administration, US)
Achim Osswald (Cologne University of Applied Sciences, DE)
Christos Papatheodorou (Ionian University, GR)
Bill Parod (Northwestern University, US)
David Pearson (National Library of Australia, AU)
Andreas Rauber (Vienna University of Technology, AT)
Seamus Ross (University of Toronto, CA)
Raivo Ruusalepp (Estonian Business Archives, EE)
Tetsuo Sakaguchi (University of Tsukuba, JP)
Lisa Schiff (California Digital Library, US)
Michael Seadle (Humbold University, DE)
Robert Sharpe (Tessela, UK)
Barbara Sierman (KB, NL)
Tobias Steinke (German National Library, DE)
Randy Stern (Harvard University, US)
Stephan Strodl (Vienna University of Technology, AT)
Shigeo Sugimoto (University of Tsukuba, JP)
David Tarrant (Southampton University, UK)
Daniel Teruggi (Institut National de l'Audiovisuel, FR)
Manfred Thaller (University of Cologne, DE)
Susan Thomas (University of Oxford, UK)
Emma Tonkin (UKOLN, University of Bath, UK)
Jeffrey Van Der Hoeven (Koninklijke Bibliotheek, NL)
Richard Wright (BBC, UK)

CONTENTS

Governance

A Capability Model for Digital Preservation: Analysing Concerns, Drivers, Constraints, Capabilities and Maturities 1
Christoph Becker, Gonçalo Antunes, José Barateiro and Ricardo Vieira

Certification and Quality: a French Experience 11
Marion Massol, Olivier Rouchon and Lorène Béchard

Users' Trust in Trusted Digital Repository Content 20
Devan Ray Donaldson

Migration

Evaluation of a Large Migration Project 24
Alex Thirifays, Anders Bo Nielsen and Barbara Dokkedal

Developing a Robust Migration Workflow for Preserving and Curating Hand-Held Media 33
Angela Dappert, Andrew N. Jackson and Akiko Kimura

Towards an Integrated Media Transfer Environment: a Comparative Summary of Available Transfer Tools and Recommendations for the Development of a Toolset for the Preservation of Complex Digital Objects 44
Antonio Ciuffreda, David Anderson, Janet Delve, Leo Konstantelos, Dan Pinchbeck, Winfried Bergmeyer, Andreas Lange and Vincent Joguin

Risk and Planning

Impact Assessment of Decision Criteria in Preservation Planning 52
Markus Hamm and Christoph Becker

Simulating the Effect of Preservation Actions on Repository Evolution 62
Christian Weihs and Andreas Rauber

Risk Assessment in Digital Preservation of e-Science Data and Processes 70
Sara Canteiro and José Barateiro

Infrastructures

Using Grid Federations for Digital Preservation 81
Gonçalo Antunes and Helder Pina

Using Automated Dependency Analysis To Generate Representation Information 89
Andrew N. Jackson

Cyberinfrastructure Supporting Evolving Data Collections 93
Maria Esteva, Christopher Jordan, Tomislav Urban and David Walling

Cost Models

A Cost Model for Small Scale Automated Digital Preservation Archives 97
Stephan Strodl and Andreas Rauber

Cost Aspects of Ingest and Normalization 107
Ulla Bøgvad Kejser, Anders Bo Nielsen and Alex Thirifays

The Costs and Economics of Preservation 116
Neil Grindley

Complex Scenarios

Long-Term Sustainability of Spatial Data Infrastructures: A Metadata Framework and Principles of Geo-Archiving 120
Arif Shaon, Carsten Rönsdorf, Urs Gerber, Kai Naumann, Paul Mason, Andrew Woolf, Michael Kirstein, Marguérite Bos and Göran Samuelsson

Short Term Preservation for Software Industry 130
Daniel Draws, Sven Euteneuer, Daniel Simon, and Frank Simon

New Dimension in Relational Database Preservation: Rising the Abstraction Level 140
Ricardo André Pereira Freitas and José Carlos Ramalho

Emulation

Replicating Installed Application and Information Environments onto Emulated or Virtualized Hardware 148
Dirk von Suchodoletz and Euan Cochrane

Remote Emulation for Migration Services in a Distributed Preservation Framework 158
Dirk von Suchodoletz, Klaus Rechert and Isgandar Valizada

Emulation as a Business Solution: the Emulation Framework 167
Bram Lohman, Bart Kiers, David Michel and Jeffrey van der Hoeven

Design Decisions in Emulator Construction: A Case Study on Home Computer Software Preservation 171
Mark Guttenbrunner and Andreas Rauber

Developing Virtual CD-ROM Collections: The Voyager Company Publications 181
Geoffrey Brown

A Braille Conversion Service Using GPU and Human Interaction by Computer Vision 190
Roman Graf and Reinhold Huber-Mörk

Cases and Communities

Evolving Domains, Problems and Solutions for Long Term Digital Preservation 194
Orit Edelstein, Michael Factor, Ross King, Thomas Risse, Eliot Salant and Philip Taylor

Recordkeeping in Temporary Command Settings 205
Erik A.M. Borglund

We are All Archivists: Encouraging Personal Digital Archiving and Citizen Archiving on a Community Scale 210
Leslie Johnston

The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data 215
Amy Pienta, George Alter and Jared Lyle

UPData - A Data Curation Experiment at U.Porto using DSpace 224
João Rocha da Silva, Cristina Ribeiro and João Correia Lopes

Towards the Preservation of Scientific Workflows 228
David De Roure, Khalid Belhajjame, Paolo Missier, José Manuel Gómez-Pérez, Raúl Palma, José Enrique Ruiz, Kristina Hettne, Marco Roos, Graham Klyne and Carole Goble

People Mashing: Agile Digital Preservation and the AQuA Project 232
Paul Wheatley, Bo Middleton, Jodie Double, Andrew N. Jackson and Rebecca McGuinness

From the World Wide Web to Digital Library Stacks: Preserving the French Web Archives 237
Clément Oury and Sébastien Peyrard

Virtual Archiving for Public Opinion Polls 242
Jonathan Crabtree

Demonstrations

Emulation Reading Room Prototype 246
Sebastian Schmelzer, Dirk von Suchodoletz and Klaus Rechert

Migration-by-Emulation 248
Isgandar Valizada, Klaus Rechert and Dirk von Suchodoletz

Re-awakening the Philips Videopac: From an Old Tape to a Vintage Feeling on a Modern Screen 250
Mark Guttenbrunner and Andreas Rauber

Meet RODA, Full-Fledged Digital Repository for Long-Term Preservation 252
Rui Castro, Luís Faria and Miguel Ferreira

Poster Abstracts

A Non-Proprietary RAID Replacement for Long Term Preservation Systems 254
Samuel Goebert and Alain Sarti

An Open-Source System for Automatic Policy-Based Collaborative Archival Replication 256
Thu-Mai Christian, Jonathan Crabtree, Nancy McGovern and Micah Altman

TOTEM: Trusted Online Technical Environment Metadata: A Long-Term Solution for a Relational Database / RDF Ontologies 258
Janet Delve, Leo Konstantelos and Antonio Ciuffreda

Corporate Recordkeeping: New Challenges for Digital Preservation 260
Gillian Oliver and Fiorella Foscarini

Preserving Change: Observations on Weblog Preservation 262
Yunhyong Kim and Seamus Ross

Transformation Rules for Model Migration in Relational Database Preservation 265
Arif Ur Rahman, Gabriel David and Cristina Ribeiro

Considerations for High Throughput Digital Preservation 267
Jason Pierson, Mark Evans, James Carr and Robert Sharpe

How Clean is Your Software? The Role of Software Validation in Digital Preservation Research Projects 269
Leo Konstantelos, Perla Innocenti and Seamus Ross

Long-Term Storage Features of Optical Disks According to Recording Conditions 271
Kwan-Yong Lee, Won-Ik Cho and Young-Joo Kim

Curation and Preservation of Research Data in Germany: A Survey Across Different Academic Disciplines 274
Achim Osswald, Heike Neuroth and Stefan Strathmann

Managing Preservation Networks: Issues of Scale for Scientific Research Assets 276
Esther Conway, Simon Lambert, Brian Matthews and Arif Shaon

The Data Management Skills Support Initiative: Synthesising Postgraduate Training in Research Data Management 279
Laura Molloy and Kellie Snow

Capitalizing on the State-of-the-Art in Preserving Complex Visual Digital Objects: The POCOS Project 282
Leo Konstantelos, David Anderson, Janet Delve, Milena Dobreva, Clive Billenness, Richard Beacham, Drew Baker, Vincent Joguin and Sonia Séfi

Building Digital Preservation Practices, Tools and Services on Quicksand 285
Bram van der Werf

Author index 287

Blank page

A Capability Model for Digital Preservation

Analyzing Concerns, Drivers, Constraints, Capabilities and Maturities

Christoph Becker
Vienna University of Technology
Vienna, Austria
becker@ifs.tuwien.ac.at

Gonçalo Antunes, José Barateiro, Ricardo Vieira
INESC-ID Information Systems Group, Lisbon, Portugal
{goncalo.antunes,jose.barateiro,rjcv}@ist.utl.pt

ABSTRACT

The last decade has seen a number of reference models and compliance criteria for Digital Preservation (DP) emerging. However, there is a lack of coherence and integration with standards and frameworks in related fields such as Information Systems; Governance, Risk and Compliance (GRC); and Organizational Engineering. DP needs to take a holistic viewpoint to acommodate the concerns of information longevity in the increasingly diverse scenarios in which DP needs to be addressed. In addition to compliance criteria, maturity models are needed to support focused assessment and targeted process improvement efforts in organizations. To enable this holistic perspective, this article discusses the question of capability maturity and presents a capability model for DP. We further demonstrate how such an architectural approach can be used as a basis to analyze the impact of criteria and metrics from the ISO Repository Audit and Certification standard on stakeholders, concerns, drivers, goals, and capabilities. The analysis presented here shall contribute to advance the understanding of cross-cutting concerns and the dicussion on maturity models in DP.

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, Design, Standardization

Keywords

OAIS Model, Repository Audit and Certification, Trust, Digital Preservation, Reference Architecture, Standards

1. INTRODUCTION

The last decade has seen considerable progress in clarifying the boundaries, goals and reference frameworks of DP. However, the relationships with related key disciplines such as Information Systems and Information Technology Management are still unclear. DP was originally driven strongly by the cultural heritage sector. Yet today, it is relevant for organizations in increasingly diverse business domains, ranging from the pharmaceutical sector to eScience and potentially any domain where information plays a key role. DP in an information-centric scenario is a cross-cutting capability orthogonal to the value chain. It has been increasingly found of fundamental importance for enabling the actual value delivery of organizations outside the traditional memory sector. DP operations are support functions to organizations that manage information and often intersect with information, services and technology across entire enterprises.

In the domain of DP, reference models for archival systems and corresponding compliance criteria have been developed. However, the general perspectives of fields such as Enterprise Architecture, Information Systems, and Governance, Risk and Compliance have not yet been fully considered. This poses a substantial barrier to increasing the recognition of DP in the mainstream fields of Information Systems and Information Technology. Furthermore, it has the effect that research in DP is often neglecting the conceptual models and powerful design techniques in fields such as Organizational Engineering, Software Engineering, and Information Systems.

The ISO 16363 standard is refining compliance criteria for repositories based on the OAIS Reference Model. The risk assessment method DRAMBORA[1] provides a catalogue of typical risks in DP environments [22]. These standards were developed specifically for traditional DP scenarios. Their focus on providing a system to address the DP problem as a whole makes it difficult to apply them in non-traditional DP settings. They deliver some guidance on compliance criteria to be met, but do not provide effective mechanisms for governance and control, or clear guidance on how to improve the processes of an organization with particular consideration of DP concerns. However, DP is becoming increasingly a concern in non-traditional environments, where the organizational environment may not be well suited for employing a DP *system* such as an OAIS-based approach, but instead requires an incorporation of DP abilities into the organizational and technological system, alongside existing processes and capabilities.

In this paper, we present a *capability model* for digital preservation that is based on established architectural principles and frameworks. We analyze this capability model

[1] `http://www.repositoryaudit.eu/`

from two perspectives. First, we discuss a *capability maturity model* based on CMMI and a method of assessing capability maturity for operational preservation. Second, we discuss the impact that criteria for trustworthy repositories as defined in ISO 16363 have on specific capabilities. The analysis presented shall be contributing to a clarification of maturity models in the field as well as an improved understanding in the implications that regulatory constraints, business drivers, and organizational goals have on organizational processes in the domain of DP.

This paper is structured as follows. Section 2 outlines related approaches and standards in the areas of DP, GRC, and Enterprise Architecture. Section 3 presents a capability model relating stakeholders and their concerns to drivers and constraints, goals, and capabilities. Section 4 discusses a maturity model for preservation operations. Section 5 relates the capability model to criteria for trustworthy repositories and illustrates the possibilities for analysis in organizational environments on a case study. Finally, Section 6 draws conclusions and gives an outlook on current and future work.

2. RELATED WORK

Digital preservation is a problem with many facets. It essentially surfaces in any organization that has to manage information over time. However, initiatives on digital preservation have been strongly driven by memory institutions and the cultural heritage sector [31]. The OAIS Reference Model [16] describes an information model and a conceptual model of key functional entities. It includes a high-level contextual view of an archival organization and its key stakeholders, and has provided a common language for the domain. However, it is difficult to reconcile these views with scenarios where different systems are in place, where related concerns may overlap with DP concerns and processes. This may for example occur in organizations where an Electronic Records Management System or an Enterprise Content Management System is in place. Key models in Records Management are the "Model Requirements for Records Systems' (MoReq2010) [12] and ISO 15489 [17]. Moreq2010 specifies functional requirements for an Electronic Records Management System and covers wide spectrum of aspects in hundreds of requirements statements. The Preservation Metadata Implementation Strategies (PREMIS) working group maintains a data dictionary for DP that contains intellectual entities, objects, rights, events, and agents [26] in a technically neutral model.

The 'Trusted Digital Repositories: Attributes and Responsibilities' report [27] (TDR) was a key milestone towards the standardization of criteria catalogs for trustworthy repositories. With the goal of providing audit and certification facilities, the Repositories Audit and Certification Criteria (RAC) are currently undergoing ISO standardization. They describe criteria for trustworthiness in the areas of Organizational Infrastructure; Digital Object Management; and Technologies, Technical Infrastructure, and Security [10, 19].

While these reference models deliver some guidance on compliance criteria to be met, they do not describe effective mechanisms for governance and control nor guidelines on implementation and improvement. However, they describe typical stakeholders and their goals and interests; recurring regulatory drivers and constraints; contractual structures, roles, and interaction patterns; solution practices and building blocks; and value propositions. As such, they are invaluable sources of domain knowledge.

DP problems, systems, and organizational concerns require a holistic, integrated view that combines aspects of organizational processes, contextual concerns, regulatory compliance and IT with systemic approaches for governance and control. These viewpoints are a stronghold of Enterprise Architecture (EA). The discipline of EA models the role of information systems and technology on organizations in a system architecture approach [15] in order to align enterprise-wide concepts, business processes and information with information technology and information systems. The core driver is planning for change and providing self-awareness to the organization in a holistic way [29]. The Zachman framework is a very influential early EA approach [32]. It describes the elements of an enterprise's systems architecture in a table where each cell is related to the set of models, principles, services and standards needed to address a specific concern of a specific stakeholder. The leading EA frameworks today are The Open Group Architecture Framework (TOGAF) [29] and the Department of Defense Architecture Framework (DODAF) [11].

IT Governance focuses on "the leadership, organisational structures and processes that ensure that the enterprise's IT sustains and extends the organisation's strategies and objectives" [8]. A widely known framework is COBIT: Control Objectives for IT. It provides a thoroughly defined process model linking resources, activities, processes and goals. One of the core concepts in Governance and Process Improvement is the idea of process *maturity*. It has been demonstrated that formal maturity models such as the Capability Maturity Model Integration (CMMI) are powerful tools for targeted improvement of processes based on quantitative assessment [14]. COBIT states that "... maturity modeling enables gaps in capabilities to be identified and demonstrated to management. Action plans can then be developed to bring these processes up to the desired capability target level" [8]. These target levels are defined in correspondence to the Software Engineering Institute's CMMI [7, 14] as (0) Non-existent, (1) Initial/Ad-Hoc, (2) Repeatable but Intuitive, (3) Defined, (4) Managed and Measurable, and (5) Optimized [8]. The maturity of processes is analyzed in the capability dimension, but not in the coverage and control dimensions. However, COBIT provides powerful controls for measuring processes both internally and externally through process and activity metrics and goal fulfillment. These concepts can be leveraged for preservation processes [3].

A recent analysis in the DP domain applied IBM's Component Business Model approach to relate DP-related business components to business areas with common objectives and evaluated the alignment of organizational structures with changing requirements of collections management and digital preservation [30]. The first SHAMAN Reference Architecture (SHAMAN-RA) presented in [2] has strong foundations in EA. However, it does not explicitly take existing domain knowledge and reference models into account in a degree sufficient to enable their transparent convergence. Based on these observations, recent work accommodated and explicitly expressed DP domain knowledge in the framework of an established Enterprise Architecture approach [1] and integrated DP capabilities with IT Governance [3]. The work presented here advances this by introducing a detailed capability model for preservation capabilities, specifying ca-

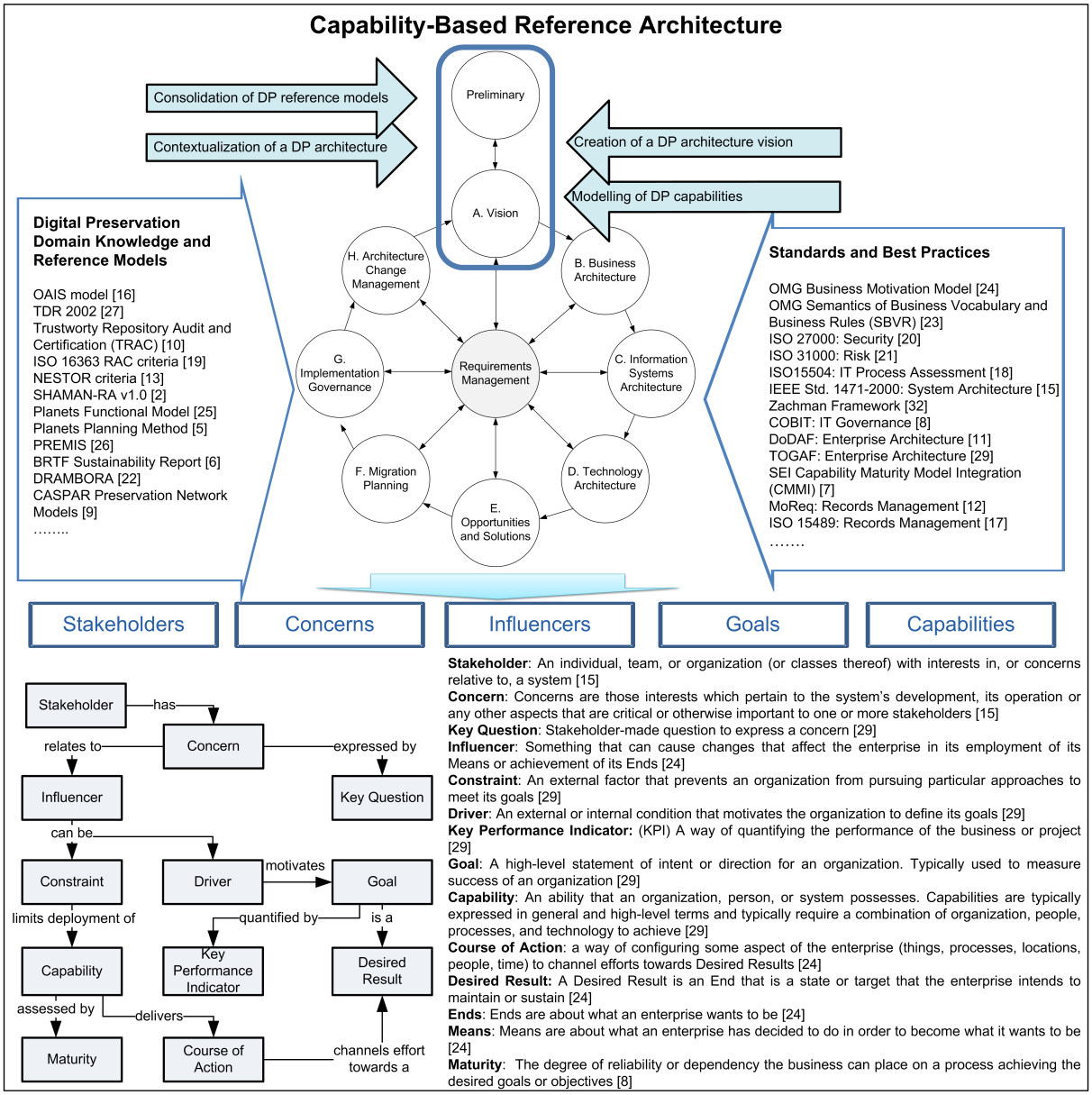


Figure 1: Using TOGAF to blend Reference Models into the SHAMAN Reference Architecture.

pability maturities for operative preservation, and analyzing the relationships between compliance criteria, drivers, stakeholders' concerns, and capabilities.

3. DIGITAL PRESERVATION CAPABILITIES

The main goal of a *Reference Architecture* is to provide a process from which concrete architecture artifacts can be derived [2]. The architecture described in [1, 3] is strongly based on TOGAF and combines it with key concepts of the Business Motivation Model (BMM) [24]. It is centered on the concept of capabilities. Note that a *capability* is fundamentally different from a system function or a process. It is instead viewed as a goal-oriented concept. A capability in TOGAF is an 'ability that an organization, person, or system possesses. Capabilities are typically expressed in general and high-level terms and typically require a combination of organization, people, processes, and technology to achieve' [29].

A successful architecture has to correctly reflect the concerns of the stakeholders of the system, from end users to developers, providing answers to whatever pertinent questions they might have. Typical digital preservation concerns include diverse aspects such as system end-usage, management, compliance, operations, and solutions.

A capability model for DP needs to be inherently independent of the business domain and, in particular, independent of the organizational scenario in which DP is deployed. It should be applicable equally to a traditional archival organization as to a business organization that is adding DP as a support capability to its primary business capabilities. It should further support organizations in answering critical questions such as *'What is the impact of a certain regulatory constraint? How can it be addressed?'* and *'How can*

we assess our processes and abilities against best practices? How can we develop targeted strategies for improvement?'.

The TOGAF Architecture Development Method (ADM) is the core component of TOGAF. It provides a systematic framework for developing an enterprise architecture. It is centered around requirements management and provides a continuous process for addressing contextual concerns and changing requirements to ensure the organization's business and IT needs are met.

We leverage the ADM to accommodate domain-specific concerns represented in DP knowledge bases and reference models. Following the ADM's first two phases, *Preliminary* and *Architecture Vision*, this requires a number of analytical steps to consolidate DP reference models, contextualize a DP architecture, model DP capabilities, and create a DP architecture vision [1].

Figure 1 illustrates the key elements of the Reference Architecture. The cyclic ADM workflow picture in the top center serves as the catalysator process into which DP domain knowledge and reference models are fed. These provide the architecture context [1], guided by standards and best practices in areas such as Information Systems; GRC; Organizational Engineering; Enterprise Architecture; and Software Engineering. Additional sources were considered, but space constraints prevents a full discussion of domain knowledge sources and their representation on the diagram. The result of our analysis is a capability-based Reference Architecture for DP that relates stakeholders and their concerns to the relevant drivers and constraints, and connects this to desired goals and required capabilities. The core concepts and their definitions and relationships are given in the bottom of Figure 1. The Reference Architecture can be used to derive concrete architectures in diverse scenarios where DP is of concern. For any concrete instantiation, additional situation-specific concerns are integrated and reconciled to produce a specific architecture by relying on the ADM process model. While previous discussions of this model focused on the high-level relations between the capabilities and their integration within an organization [1, 3, 4], we will focus here on the detailed component capabilities of preservation and specify a maturity model for preservation operations. We further outline performance measures that can be used to assess the maturity and performance of organizational capabilities along a number of dimensions.

From the analysis of the DP references, several stakeholders were identified. Stakeholders with end-usage concerns include the *Producer/ Depositor* and the *Consumer* stakeholders, which are identical in definition to the OAIS *Producer* and *Consumer* roles. The *Producer/ Depositor* stakeholder is the entity responsible for the ingestion of the objects to be preserved. Typically, its concerns include: the deposit of objects along with whatever additional data required, in accordance with negotiated agreements/contracts; assurance of access rights to the objects; assurance of the authenticity of provenance of the deposited objects; and preservation of the objects and associated rights beyond the lifetime of the repository. The *Consumer* stakeholder represents users accessing the preserved objects, with a potential interest in its reuse and a certain background in terms of knowledge and technical environment. Its concerns include the access to the preserved objects in accordance with negotiated agreements/contracts, and the correspondence of the retrieved content to its needs in terms of understandability and authenticity. Other identified stakeholders include *Management*, a generalization of all management stakeholders concerned with ends and means. Specializations of the *Management* stakeholder include the *Executive Management*, *Repository Manager*, *Technology Manager*, and *Operational Manager*. Stakeholders with compliance concerns include the *Regulator* and the *Auditor*. Operational concerns are shared between the *Repository Operator* and *Technology Operator*. Finally, stakeholders with solutions-related concerns include the *System Architect* and the *Solution Provider*.

The analysis of stakeholders' concerns, typical compliance requirements, domain models and other sources of knowledge enables an analysis of the main influencers that have an impact on the setting of organizational goals in digital preservation. Such influencers can be either drivers or constraints. The key distinction made between these influencers is between *internal* and *external* influencers. These influencers in turn drive and constrain an organization's definition of high-level goals, i.e. the desired results that an organization wants to achieve. Such goals strongly relate to stakeholders' concerns such as the user community's perception of content's authenticity, and require certain abilities inside the organization to achieve corresponding outcomes. A detailed discussion and categorization of DP drivers, an assessment of possible constraints (through external drivers), and an analysis of exemplary DP goals and their associated Key Performance Indicators is described in [1].

The organization's stakeholders, concerns, and goals in turn drive the clarification of its value chain definition and, finally, the specification of the abilities that it needs to achieve its stated goals. Figure 2 shows the high-level capability model. Capabilities are grouped into governance capabilities, business capabilities and support capabilities. In general, governance capabilities control business and support capabilities; business and support capabilities inform governance capabilities; and business capabilities depend on support capabilities. These high-level capabilities are described in [4]. The core business capability of DP in this model is **Preserve Contents** – the 'ability to maintain content authentic and understandable to the defined user community over time and assure its provenance' [1]. This is at the heart of DP, it addresses the core requirement of authenticity, understandability and provenance. This core capability is composed of two capabilities: Preservation Planning and Preservation Operation. **Preservation Planning** is 'the ability to monitor, steer and control the preservation operation of content so that the goals of accessibility, authenticity, usability and understandability are met with minimal operational costs and maximal (expected) content value. This includes managing obsolescence threats at the logical level as the core risk affecting content's authenticity, usability and understandability'[3].

Preservation Planning consists (at a minimum) of the capabilities **Planning Operational Preservation** and **Monitoring**. *Planning Operational Preservation* is the ability to make drivers and goals operational, i.e. define objectives and constraints represented by decision criteria, and assess options against these criteria to deliver efficient decisions and operational plans. It is composed of a number of component capabilities:

1. *Influencers and Decision Making*: The ability to make drivers and goals operational, i.e. define objectives and constraints represented by decision criteria, and

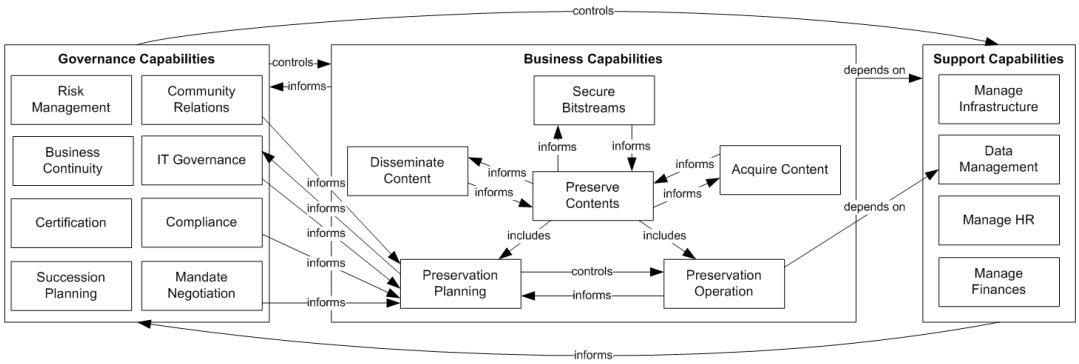


Figure 2: Capability Relations for *Preserve Contents*. Relations can be *inform, control, include, depend on.*

assess options against these criteria to deliver efficient decisions and operational plans.

2. *Options diagnosis*: The ability to gather information about available options, i.e. measures corresponding to a set of criteria.

3. *Specification and Delivery*: The ability to specify actions and directives in an understandable form and deliver it to operations (to prepare the deployment of plans).

The second planning capability, *Monitoring*, is the ability to monitor operations (in particular the execution of plans) and the environment, i.e. the ability to monitor all influencers having a potential impact on plans to ensure conformance of results to expected outcomes and notify the decision making capability (Planning Operational Preservation) of a change that requires assessment. It is composed of

1. *Internal Monitoring*: The ability to monitor operations for certain properties of interest, which include operations specified by plans and operational attributes of the system, i.e. internal influencers. Internal influencers of interest include (but are not necessarily limited to) operational statistics about the execution of delivered plans, operational statistics about content dissemination, and user feedback about satisfaction with respect to user access requirements.

2. *External Monitoring*: The ability to monitor external influencers of interest through the corresponding governance capabilities. External influencers include (but are not necessarily limited to): Technological opportunities for improving achievement of goals (through IT Governance); Technology correspondence (as an indicator of impending obsolescence, misalignment to user access requirements) (through IT Governance, Disseminate Content, Acquire Content); User community shifts (through Community Relations); Access requirements (through Disseminate Content); and Regulations (through Mandate Negotiation, Compliance).

Preservation Operation is 'the ability to control the deployment and execution of preservation plans. This includes analysing content, executing preservation actions and ensure adequate levels of provenance, handling preservation metadata, conducting Quality Assurance, and providing reports and statistics, all according to preservation plans. Preservation Actions are concrete actions (usually implemented by a software tool) performed on content in order to achieve preservation goals. For example, a migration of content to a different format using a certain tool in a certain configuration and environment'[3]. Preservation Operation is in turn composed of six component capabilities:

1. *Analysis:* The ability to measure properties of relevance in the content and document them in understandable form.

2. *Action*: The ability to execute preservation actions in order to actively preserve content according to preservation plans

3. *Quality Assurance*: The ability to deliver accurate measures that quantify the equivalence of performances (renderings) of preserved content by measuring properties of renderings/performances and comparing them to each other to measure their equivalence corresponding to requirements.

4. *Preservation Metadata*: The ability to read, understand and write appropriate preservation metadata corresponding to chosen standards.

5. *Plan Deployment*: The ability to receive plans from Planning and deploy them to an operational environment.[2]

6. *Reporting and Statistics*: The ability to produce documentation of activities in an adequate and understandable form (for monitoring and auditing).

Each of these component capabilities can be measured along a number of metrics. *Options Diagnosis*, for instance, can be measured along the following dimensions:

- Completeness: Measures are delivered for all options and each criterion.
- Correctness: All measures are correct.
- Timeliness: All measures are delivered in a certain time frame.

Similarly, the Monitoring capabilities can be tracked for completeness, correctness, timeliness and currentness.

On the operations side, performance indicators for *Actions* will include

- Completeness: Successful execution of all actions deployed as part of a plan.

[2]Technically, this may result in a set of operations potentially combining analysis, actions, QA, metadata, and reporting, all of which may be specified by the plan. Execution of the plan may require a combination of services, orchestration, and processes involving human intervention.

- Provenance: Delivery of complete audit trails to ensure provenance for every action executed.
- Results documentation: Delivery of complete information about the correspondence of action results to expected results.
- Operations documentation: Delivery of complete information about the state of operations at any point in time.

Metrics for *Reporting and Statistics* will generally include the following.

- Timeliness: Reports and statistics are delivered within a certain time frame after requested.
- Currentness: Reports and statistics always show up-to-date information, i.e. delay is below certain threshold.
- Completeness: Reports and statistics contain all relevant information about all operations.
- Relevance: Reports and statistics contain minimum unnecessary information.
- Correctness: Information reported is correct.
- Understandability: Reports and statistics are understandable by all consuming entities.

Clearly, the exact metrics that are available and meaningful in a concrete environment will depend on the organizational processes and tools available. Furthermore, the metrics described above are oriented towards an internal measurement of capabilities, and as such need to be complemented by external measures related to goal achievement. For example, the core goal of delivering authentic, understandable, and usable content to the user community can be associated with a KPI such as 'Percentage of transformational object properties preserved by actions as denoted by user feedback and/or QA measures in comparison to guarantees provided by specified SLAs' [1]. A specification of the relationships between these process metrics and the associated outcomes of capabilities measured in KPIs is needed to achieve full control over preservation processes. However, apart from goal achievement and process metrics, capabilities can also be analyzed on a more abstract level for their maturity.

4. A MATURITY MODEL FOR PRESERVATION OPERATIONS

Focusing on strategic process and capability improvement rather than formal certification of processes, COBIT provides maturity level specifications for each process along a number of dimensions similar to [18]. We can thus assess the maturity of the *Preservation Operation* capability on the dimensions (1) *Awareness and Communication*, (2) *Policies, Plans and Procedures*, (3) *Tools and Automation*, (4) *Skills and Expertise*, (5) *Responsibility and Accountability*, and (6) *Goal Setting and Measurement*.

Table 1 defines criteria for the *Preservation Operation* capability for each maturity level and dimension. Similar criteria have been specified for Preservation Planning elsewhere [3]. As an illustrative example, consider an organization with the following diagnosis on their preservation operations: *Management is aware of the role of operations for authenticity and provenance, and there is a defined process for operations. This process includes all activities (actions, analysis, Quality Assurance, Metadata, and Reporting), and it relies on standardized plans. These plans are generally deployed according to specifications, but the deployment and operation is a mostly manual process of initiating operations as far as they are concretely specified by these plans. QA and metadata management is not driven by plans, and it does not seem to be aligned with business goals. There are guidelines about statistics and reporting procedures, but no integrated system exists for tracking the state of operations and the results of actions, and no formal metrics have been defined. Several automated tools are employed in different processes. However, the processes and rules used are defined by the availability of components and services and the level of skills of the people running these processes. A formal training plan has been developed that defines roles and skills for the different sets of operations, but all training is in fact still based on individual initiatives and not continuously managed.*

Assessing the organization's capability along the dimensions outlined above, it can be considered to be on the *Defined* level for all dimensions. Considering the skills and expertise set in the example above, we can verify that staff has operational skills and a formal training plan was developed. The absence of formal responsibility and accountability plans, however, increases the organization's dependency on specific people, which increases the severity of losing key staff trained on individual initiatives and not continuously managed. Notice that in reality, processes will generally be on different maturity levels for varying dimensions [3]. Awareness and Communication, for example, often precedes automation and tool support.

This type of capability assessment provides an internal benchmarking of the quality of processes in several dimensions. The analysis provides organizations with a decision support mechanism to prioritize actions to improve the quality of their capabilities (what and how can be improved). On the other hand, we must recognize the existence of dependencies between distinct capabilities, as shown in Figure 2. For instance, Preservation Operation *informs* Preservation Planning, but depends on other capabilities. Thus, to systematically improve the performance and maturity level of specific capabilities, we also need to consider the quality of related capabilities and understand the dependencies and the relations between internal process metrics and external outcome indicators.

5. ANALYZING CONSTRAINTS, GOALS AND CAPABILITIES

When an organization intends to analyze the impact of policies, external influencers and regulatory compliance constraints, it is often unclear which areas are concerned, and how to represent the impact (and measure the fulfillment) of certain influencers. In particular the interplay between drivers and constraints and their accumulated impact on required processes and functions is difficult to assess.

A core strength of an EA-based approach is the clear definition and separation of concerns and the traceability that it provides for impact assessment of changes. Arising constraints and drivers can be assessed with respect to the effects that they cause on concerns, goals and capabilities. Relying on the conceptual model outlined above, these can thus be addressed along the following dimensions.

	Awareness and Communication	Policies, Plans and Procedures	Tools and Automation	Skills and Expertise	Responsibility and Accountability	Goal Setting and Measurement
1	Management recognizes the need for preservation operations. There is inconsistent and sporadic communication.	Some operations are carried out, but they are not controlled. No useful documentation is produced about procedures and actions.	Some tools may be employed by individuals in an unsystematic ad-hoc manner.	There is no common awareness of which skills and expertise are required for which tasks.	There is no common awareness of responsibilities.	There is no clear awareness of goals; operations solely react to incidents and are not tracked.
2	Management is aware of the role of operations for authenticity and provenance. No formal reporting process exists, but there is some documentation about process results. Reports are delivered by individuals.	Some operational procedures emerge, but they are informal and intuitive. Operations rely on individuals; different procedures are followed within the organization. QA is recognized as a process, but mostly carried out ad-hoc and manual.	Automated tools are beginning to be employed by individuals based on arising needs and availability. Their usage is unsystematic and incoherent.	Staff obtain their operational skills through hands-on experience, repeated application of techniques and informal training by their peers.	Responsibility for operations emerges, but is not documented. Accountability is not defined.	There is individual awareness of short-term goals to achieve in operations, but no consistent goal definition or measurement.
3	Management understands the role of operations for authenticity and provenance. There are guidelines about statistics and reporting procedures, but they are not consistently enforced.	There is a defined process for all operations that relies on standardized plans. The processes and rules used are defined by available components, services and skills. QA and metadata management are not driven by business goals.	Plans are deployed according to specifications, but the process of initiating operations is mostly manual. No integrated system exists for tracking the state and results of operations.	A formal training plan has been developed that defines roles and skills for the different sets of operations, but formalized training is still based on individual initiatives.	Responsibility for operations is assigned, but accountability is not provided for all operations.	Operational goals are specified, but no formal metrics are defined. Measurements take place, but are not aligned to goals. Assessment of goal achievement is subjective and inconsistent.
4	Management fully understands the role of operations for authenticity and provenance and how they relate to business goals in the organization. Reporting processes are fully specified and adhered to.	Plans are fully deployed as operational activities, and the compliance of all operations to goals and constraints specified in plans is fully monitored. All Operations are actively monitoring state of operations.	An automated system exists to control automated operations, and automated components are widespread, yet not fully integrated.	Required skills and expertise are defined for all roles, and formal training is in place.	Responsibility and accountability for all operations is clearly defined and enforced.	A measurement system is in place and metrics are aligned with goals. Compliance monitoring is supported and compliance enforced in all operations.
5	Operations are continuously improving. An integrated communication and reporting system is fully transparent and operates in real time.	Extensive use is being made of industry good practices in plan deployment, analysis, actions, metadata, QA, and reporting.	All operations are fully integrated, status is constantly available in real-time.	Operators have the expertise, skills and means to conduct all operations. Continuous skills and expertise assessment ensures systematic improvement.	A formal responsibility and accountability plan is fully traceable to all operations.	Compliance is constantly measured automatically on all levels. Continuous assessment drives the optimization of measurement techniques.

Levels: 1: Initial/Ad-Hoc, 2: Repeatable but Intuitive, 3: Defined, 4: Managed and Measurable, 5: Optimized [8]

Table 1: Maturity Levels for the capability *Preservation Operation*

- *Stakeholders concerned*: Which are the stakeholders whose interests and viewpoints are affected by the influencer? How will it change their view of the world?
- *Concerns addressed*: Which concerns will need to consider the exact implications of the influencer? Do the Key Questions accurately reflect these considerations? Is it possible to model the influencer and its impact in the defined viewpoints and perspectives that represent the concerns?
- *Drivers involved*: Which organizational drivers are involved? What is the combined effect of a regulatory constraint and a business driver on the organizational goals?
- *Goals impacted*: Which organizational goals may be affected by an influencer, and how?
- *Capabilities affected*: Which capabilities will need to consider the effect of the influencer in order to be successfully achieving their stated goals? How can they accommodate this influencer?
- *Metrics applicable*: Which Key Performance Indicators need to be tracked to detect the exact effect of an influencer on the organization's achievement of goals? Which metrics can be used to assess capabilities? How mature are our capabilities?

Consider the case of RAC 4.1.1, *The repository shall identify the Content Information and the Information Properties that the repository will preserve.* This is part of 4.1 *Ingest: Acquisition of Content.* Based on the capability-centered Reference Architecture, it becomes possible to analyze the impact of a regulatory or organizational constraint along the lines outlined above:

- *Stakeholders concerned*: The primary stakeholders concerned include Producer/ Depositor; Consumer; and Management. However, the Repository Operator and the Solution Provider may be involved, depending on the organization's process model and the decisions taken by Management.
- *Concerns addressed*: Focusing on the OAIS-related stakeholders, the concerns addressed include (Key Questions in brackets):
  1. *Producer/Depositor: Authenticity and Provenance.* Content provided is authentic and has complete provenance. (What kinds of guarantees will the repository provide to assure me the authenticity and understandability of my objects? Will complete provenance information be provided with the disseminated content, so that the provided objects be traceable to the original?)
  2. *Consumer: Content.* The information retrieved is authentic, understandable and corresponds to my needs. (Will the domain knowledge that I have be sufficient to access and understand the content? Will the objects be corresponding to my queries, authentic, compatible to my technical environment, and understandable?)
  3. *Management: Mandate, Mission, Policies and Compliance.* The governance of the mandate, the commitment of the organization to digital preservation, may it be for business needs, legal, or legislative reasons; and corresponding compliance. This includes certification and succession planning. (Is the mandate adequate, well-specified and appropriately accessible? Is the organization able to fulfill the mandate? Does the organization possess all the required contracts regarding succession planning and escrow agreements? Is the organization compliant to external regulations?

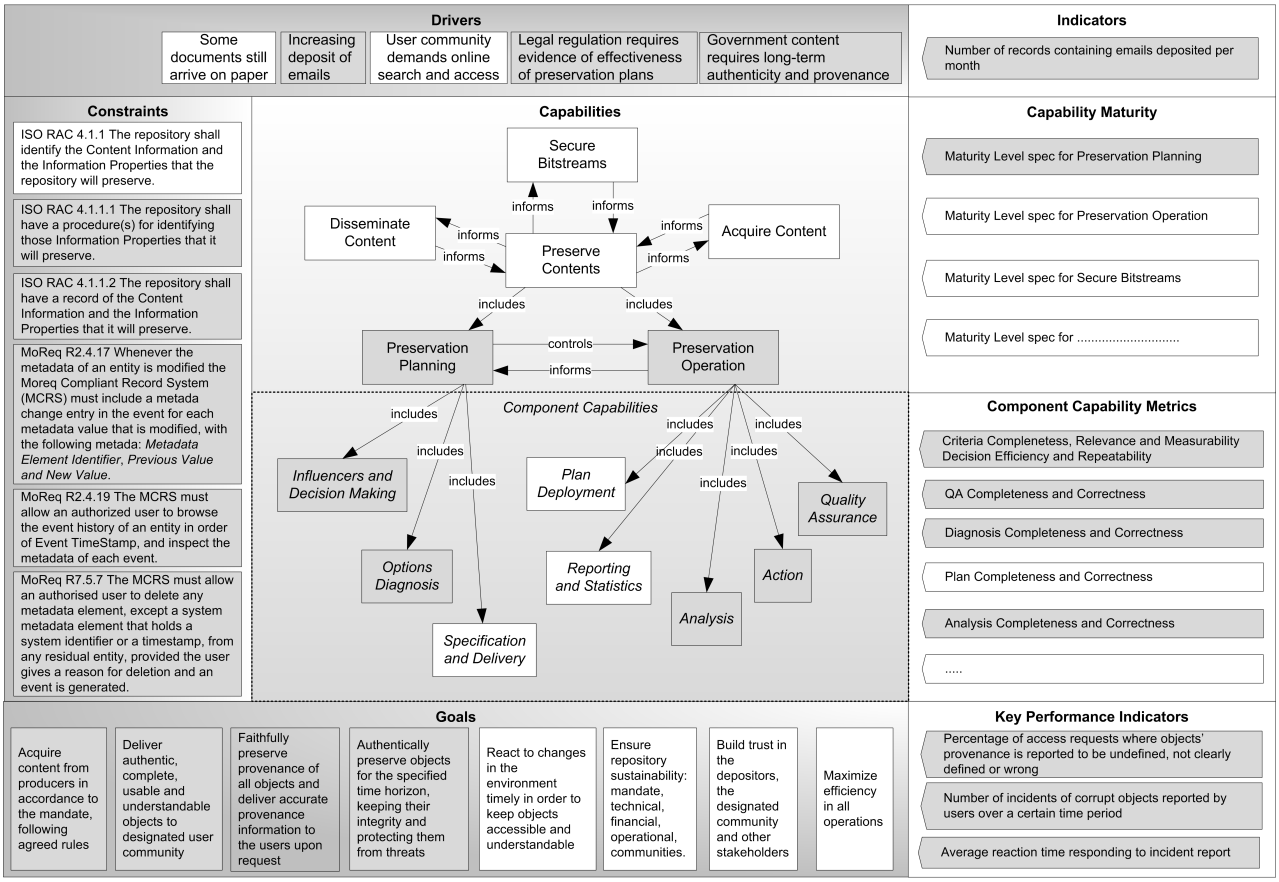


Figure 3: Business drivers and compliance to ISO RAC and MoReq2010: Content Acquisition at the CML.

Does the organization possess necessary certifications?)

- *Drivers involved*: The external drivers and constraints involved include (at least) access contracts; deposit contracts; supplier contracts; and the user community's knowledge, technology, and demand satisfaction. The internal drivers include organizational capabilities, resources (staff expertise and qualifications, existing software and costs) and the business vision.

- *Goals impacted*: The combined impact of these on the goals depends on the business vision, but will have at least an impact on the targeted level of fulfillment in terms of authenticity.

- *Capabilities affected*: Correspondingly, a number of capabilities will be affected: For example, the capability *Community Relations* may be required to consult and negotiate with communities about levels of information properties preserved. Similarly, *Preservation Planning* will need an understanding of the information properties to preserve, a reliable method for assessing the fulfillment of the goals derived from these, a method for evaluating potential ways of preserving all properties, and the ability to specify them for operational purposes. *Preservation Operation*, in turn, will need appropriate means for Analysis, Actions, and Quality Assurance that are aligned with the content that the archive has to deal with.

- *Metrics applicable*: Finally, the metrics that can be tracked can be deduced from the component capabilities affected. For example, they will include the metrics mentioned in Section 3, such as completeness and correctness of Options Diagnosis. Furthermore, external KPIs can be used to measure outcomes, and capabilities can be assessed for their maturity levels. All of these metrics can be used to assess compliance to the original regulatory constraint as specified in RAC, and used as targets to improve organizational capabilities.

As a simplified visual illustration, Figure 3 shows a real-world case where RAC 4.1.1 intersects with a business driver. The municipality of Lisbon (CML) is in the process of integrating the software Documentum[3] with a set of business workflows for a wide range of organizational entities, including the Municipal Archives. In this process, Records Management concerns overlap with DP concerns and a number of specific drivers of organizational change. Generic high-level DP goals as outlined in [1] are pictured at the bottom. Relevant business drivers to be addressed are shown on the top. These include long-term authenticity and provenance, as well as a need for evidence-based proof of effectiveness. Selected constraints posed by RAC and Moreq2010 are listed on the left. The right side shows indicators that can be tracked externally and internally to exercise control based on a quantative assessment. The related constraints, drivers, capabilities, maturities, capability metrics, goals, and KPIs are shaded in gray.

[3] `http://www.emc.com/domains/documentum/index.htm`

	Capabilities	4.1 Ingest: Acquisition of Content	4.2 Ingest: Creation of the AIP	4.3 Preservation Planning	4.4 AIP Preservation	4.5 Information Management	4.6 Access Management
Governance	Compliance	A	A	A	A	A	A
	Community Relations	S	S	S		S	
	Certification						
	Mandate Negotiation			A			
	Business Continuity						
	Succession Planning						
	IT Governance	A	A	A	A	A	A
	Manage Risks	A	A	A	A	A	A
Business	Acquire Content	R	R				
	Secure Bitstreams	S	S		S		
	Preserve Content						
	- Preservation Planning	S	S	R	S		
	- Preservation Operation	S	S		R		S
	Disseminate Content					S	R
Support	Data Management		S		S	S	
	Manage Infrastructure						
	Manage HR						
	Manage Finances						

Table 2: High-level capabilities (A)ware of, (R)esponsible for or (S)upporting RAC criteria in group 4

The increasing move towards email deposit intersects with RAC 4.1.1, since the significant properties that will be preserved need to be decided. This is a typical task for preservation planning, which will require a clear documentation of decision factors and the ability to diagnose possible options for email preservation to decide on a feasibility and level of authenticity that can be guaranteed. On an operational level, this requires processes and tools for email analysis and quality assurance for potential preservation actions. The affected component capabilities can be assessed along the measures outlined above, while Preservation Planning and Preservation Operation can be assessed for capability maturity. On the level of end-user results, i.e. business outcomes, Key Performance Indicators can be used to track goal achievement from an external perspective.

Table 2 summarizes the impact of each group of RAC criteria in section 4 (Digital Object Management) on the capabilities. Essentially, a criterion can be (part of) the primary *responsibility* of a capability, or a capability may be indirectly required to support the fulfilment. For example, operational verification of content integrity as requested in RAC 4.2 – which is primarily concerned with Ingest – requires fixity checks, which are part of Data Management. Finally, certain capabilities may need to be *aware* of compliance criteria to be successful in *their* mission. For example, *Compliance* is affected by all constraints – since its mission is to 'verify the compliance of operations and report deviations' [4], it will need to be aware of all compliance constraints. This applies a priori to Governance, Risk and Compliance, but is also required in other areas. In this sense, it is interesting to see how certain groups of criteria have an impact beyond the obvious one that refers to the directly responsible capability. For example, the criteria listed in section 4.1 influence not only the *Acquire Content* capability, but also others, such as the business capability *Secure Bitstream*. This is caused by *'4.1.6 The repository shall obtain sufficient control over the Digital Objects to preserve them.'*[19], which makes direct references to bitstream preservation.

The compliance with RAC criteria will also have an impact on the maturity level of capabilities. However, this is dependent on the way that compliance is achieved. For instance, RAC 4.1.5, *The repository shall have an ingest process which verifies each SIP for completeness and correctness*, may influence the maturity levels for the *Policies, Plans and Procedures* dimension of the Acquire Content capability. Depending on the way compliance is monitored, it can also impact the *Tools and Automation* dimension, if the verification is automated. Other dimensions will be impacted as well, although indirectly.

6. DISCUSSION AND OUTLOOK

The Reference Architecture that forms the basis of this article is an Enterprise Architecture-based approach that enables the accommodation of digital preservation concerns in the overall architecture of an organization. For that, a capability-based model of preservation was derived from established digital preservation key references and best practices from related fields. This included in-depth analysis of the stakeholders of the domain, their concerns, goals, and influencers (drivers and constraints). The result is a multidimensional view on the domain concepts covered in these key references. The approach taken with this Reference Architecture enables the transfer of DP know-how into a nontraditional repository-based DP scenario, since it is itself agnostic to concrete scenarios. In other words, this capability-based approach can deliver value to organizations in which the preservation of contents is not a main business requirement, but required to enable actual delivery of value in the primary business.

The specification of internal process metrics and external metrics measuring the achievement of certain goals by each capability through KPIs represents an essential step towards a quantified control mechanism that can be used effectively to exercise control and govern capabilities [3].

The approach provides a powerful tool to enable responsible stakeholders to analyze the impact of compliance regulations and constraints on their systems' architecture requirements and their organizational capabilities. It can furthermore be used to assess capability maturity and process maturity to enable focused improvement of key areas. It thus enables organizations to improve maturities by considering the impact that compliance requirements have on organizations' capabilities and processes. Based on a maturity assessment, an organization can target a *capability increment* to improve its capabilities and their maturities by undergoing a change initiative to increase performance for a particular capability [29].

Current work is focused on moving forward in the TOGAF-

ADM cycle to derive a contextualized *Business Architecture* for a concrete real-world scenario, and conducting a full-depth analysis of the combined implications of constraints coming from the domains of DP and Records Management in a real-world case. This furthermore sets the grounds for a full maturity model on all capabilities.

Acknowledgments

This work was supported by FCT (INESC-ID multi-annual funding) through the PIDDAC Program funds and by the projects SHAMAN and SCAPE, funded under FP7 of the EU under contract 216736 and 270137, respectively.

7. REFERENCES

[1] G. Antunes, J. Barateiro, C. Becker, R. Vieira, and J. Borbinha. Modeling contextual concerns in Enterprise Architecture. In *Fifteenth IEEE International EDOC Conference*, Helsinki, Finland, August 29 - September 2 2011.

[2] G. Antunes, J. Barateiro, and J. Borbinha. A reference architecture for digital preservation. In *Proc. iPRES2010*, Vienna, Austria, 2010.

[3] C. Becker, G. Antunes, J. Barateiro, R. Vieira, and J. Borbinha. Control Objectives for DP: Digital Preservation as an Integrated Part of IT Governance. In *Proc. 74th Annual Meeting of ASIST*, New Orleans, October 2011.

[4] C. Becker, G. Antunes, J. Barateiro, R. Vieira, and J. Borbinha. Modeling digital preservation capabilities in enterprise architecture. In *In 12th Annual International Conference on Digital Government Research (dg.o 2011), June 12-15, College Park, MD, USA.*, 2011.

[5] C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman. Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. *Int. Journal on Digital Libraries (IJDL)*, December 2009.

[6] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. *Sustainable Economics for a Digital Planet*. 2010.

[7] Software Engineering Institute. Capability Maturity Model Integration for Development. Version 1.3. Carnegie Mellon University, November 2010.

[8] IT Governance Institute. CobiT 4.1. framework – control objectives – management guidelinces – maturity models, 2007.

[9] E. Conway, M. Dunckley, B. Mcilwrath, and D. Giaretta. Preservation network models: Creating stable networks of information to ensure the long term use of scientific data. In *Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data*, Villafranca del Castillo, Madrid, Spain, 2009.

[10] CRL and OCLC. Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC). Technical report, The Center for Research Libraries and Online Computer Library Center, February 2007.

[11] Department of Defense, Washington D.C. *DoD Architecture Framework, Version 2.0*, 2009.

[12] DLM Forum Foundation. *MoReq2010 - Model Requirements for Records Systems. Draft - v0.92*, 2010.

[13] S. Dobratz, A. Schoger, and S. Strathmann. The nestor catalogue of criteria for trusted digital repository evaluation and certification. In *Proc. JCDL2006*, 2006.

[14] D. L. Gibson, D. R. Goldenson, and K. Kost. *Performance Results of CMMI-Based Process Improvement*. Software Engineering Institute, Pittsburg, PA, 2006.

[15] IEEE. *Recommended Practice for Architecture Description of Software-Intensive Systems (IEEE 1471-2000)*. IEEE Computer Society, 2000.

[16] ISO 14721:2003. Open archival information system – Reference model, 2003.

[17] ISO 15489-1:2001. Information and documentation: Records management, 2001.

[18] ISO/IEC 15504-1:2004. Information technology - Process assessment – Part 1: Concepts and Vocabulary, 2004.

[19] ISO/DIS 16363. Space data and information transfer systems - Audit and certification of trustworthy digital repositories. Standard in development, 2010.

[20] ISO/IEC 27000:2009. Information technology - Security techniques - Information security management systems - Overview and Vocabulary, 2009.

[21] ISO 31000:2009. Risk management – Principles and guidelines, 2009.

[22] A. McHugh, R. Ruusalepp, S. Ross, and H. Hofman. The digital repository audit method based on risk assessment (DRAMBORA). In *Digital Curation Center and Digital Presevation Europe*, 2007.

[23] Object Management Group. *Semantics of Business Vocabulary and Business Rules (SBVR), Version 1.0*. OMG, 2008.

[24] Object Management Group. *Business Motivation Model 1.1*. OMG, May 2010.

[25] PLANETS Consortium. Report on the planets functional model. Pp7/d3-4, 2009.

[26] PREMIS Editorial Committee. *PREMIS Data Dictionary for Preservation Metadata version 2.1*, January 2011.

[27] RLG/OCLC Working Group on Digital Archive Attributes. *Trusted Digital Repositories: Attributes and Responsibilities*. Research Libraries Group, 2002.

[28] C. Rosenthal, A. Blekinge-Rasmussen, J. Hutar, A. McHugh, S. Strodl, E. Witham, and S. Ross. *Repository Planning Checklist and Guidance*. HATII at the University of Glasgow, 2008.

[29] The Open Group. *TOGAF Version 9*. Van Haren Publishing, 2009.

[30] R. J. van Diessen, B. Sierman, and C. A. Lee. Component business model for digital repositories: A framework for analysis. In *Proc. iPRES 2008*, 2008.

[31] C. Webb. *Guidelines for the Preservation of Digital Heritage*. Information Society Division United Nations Educational, Scientific and Cultural Organization (UNESCO) – National Library of Australia, 2005.

[32] J. Zachman. A framework for information systems architecture. *IBM Systems Journal*, 12(6):276–292, 1987.

Certification and Quality: A French Experience

Marion MASSOL
CINES
950, rue de Saint Priest
34097 MONTPELLIER Cedex 5
(+33) 4 67 14 14 86
massol@cines.fr

Olivier ROUCHON
CINES
950, rue de Saint Priest
34097 MONTPELLIER Cedex 5
(+33) 4 67 14 14 67
rouchon@cines.fr

Lorène BECHARD
CINES
950, rue de Saint Priest
34097 MONTPELLIER Cedex 5
(+33) 4 67 14 14 55
bechard@cines.fr

ABSTRACT
The CINES has two main missions, among which is the long-term preservation of French scientific data. To provide this service, CINES deployed in 2006 one of the first digital repository in France named PAC (Plateforme d'Archivage du CINES – the CINES preservation system).

In order to secure this mandate in the long-term, it is absolutely crucial for CINES to prove the quality of the services it provides to the French higher education and research community. For this purpose, the CINES strategy relies on the adoption of a quality assurance approach which includes the certification of its repository.

Over the past four years, the PAC staff ran not less than five audits, internal as much as external. Various systems of reference have been used: some were at the national level (National Archives accreditation), others were at a European level (Data Seal of Approval accreditation, DRAMBORA) or even at an international level (ISO 16 363, TRAC).

From these audits, the strengths and weaknesses of the digital preservation repository have been highlighted. Action plans have been put together and executed to improve the service quality. The aim of transparency, which ranked first in the certification initiative, also reinforced the trust of the user community toward the long term digital preservation service of the CINES. Based on such an experience, the PAC staff is now willing to share its knowledge and feedback with the rest of the community, by participating in think tanks as well as standardization workgroups.

Categories and Subject Descriptors
H.3.7 **[Information storage and retrieval]**: Digital Libraries – *Standards, Systems issues.*

K.6.4 **[Management of Computing and Information systems]**: System Management - *Management audit, Quality assurance.*

K.7.3 **[The Computing Profession]**: Testing, Certification, and Licensing.

iPRES2011, Nov. 1–4, 2011, Singapore.

General Terms
Standardization, Measurement, Documentation, Verification

Keywords
Certification, Audit, Quality, Trust, Long-term Preservation, Archive, Risk Management, Metrics and Assessment.

1. BACKGROUND
1.1 CINES and digital preservation overview
CINES (Centre Informatique National de l'Enseignement Supérieur) is a French IT datacenter for the Higher Education and Research community. This state administration institution employs about 50 engineers, is based in Montpellier and is known worldwide for its HPC (high performance computing) activities. The whole of the CINES means is made available for all the French researchers, who are gathered together in scientific domains. The largest communities to use the CINES computing infrastructure are the fluid mechanics, chemistry and climatology research communities.

As part of this first mission, CINES hosts advanced computers which include Jade (SGI ICE 8200 EX with 267 TFlops peak, 23 040 cores and 700TB of disks), the 7th most powerful supercomputer in Europe and 30th in the international TOP 500 ranking (June, 2011).

The second main activity of CINES is the long-term preservation of records and data with one of the very few operational long-term preservation platforms in France. This archiving repository is called PAC [3] (Plateforme d'Archivage du CINES – the CINES preservation system).

The very first thoughts on digital preservation were given in 2004. In 2006, the first digital PhD theses were archived on PAC-V1 (which was developed internally). Starting march 2008, the documents are preserved on PAC-V2, which relies on the Arcsys software edited by Infotel and on specific additional modules (Ingest module, Archeck – data integrity control application, ArcStats – statistics tool, representation information library…) developed in-house. Four copies are made of the archives: two are kept on hard disk drives, and two are stored on a tape library.

The archival processes are fully automated. The only manual interventions are performed at the beginning of every archive project: appraisal of digital objects to be preserved, data mapping between the producer information system and the CINES metadata model, agreement on file formats, definition of the package structure, user tests, etc. Thus, project after project, the staff of the Digital Preservation Department has increased. At this stage, there are 11 people in the preservation team with different knowledge, skills and experiences. There are:

- ✓ An I/T manager;
- ✓ An archivist;
- ✓ A File formats expert (assisted by an expert on video file formats);
- ✓ I/T developers;
- ✓ System administrators;
- ✓ A XML specialist;
- ✓ Hardware and OS specialists;
- ✓ Service support and monitoring specialists (24x7).

Three types of digital documents are secured on PAC for the years to come:

- ✓ Scientific data generated from observations, measurements or computation;
- ✓ Heritage data like PhD theses, educational data or pedagogics, publications or scientific digitized books;
- ✓ Administrative data from French universities: civil servants' records…

At present, there are about 13 TB of data in the production environment:

- ✓ Digital PhD theses;
- ✓ Scientific papers uploaded in the open repository HAL (Hyper Article on Line) managed by CCSD;
- ✓ Digitized publications as part of the Humanities and Social Sciences program « Persée »;
- ✓ CRDO Multimedia collection (sound files of ethnographic recordings in various languages) as part of the Humanity and Social Sciences program « TGE-Adonis »;
- ✓ Digitized collection of the history of law of CUJAS university library;
- ✓ Digitized collection of books about the History of Medicine (BIU Santé - Inter-university library of healthcare);
- ✓ Digitized works in medicine, biology, geology and physics, chemistry (BUPMC - University Library "Pierre and Marie Curie");
- ✓ Library of photos of the French School of Far East.

CINES has other projects to preserve: "Canal U" CERIMES multimedia collection (audiovisual files of recordings of courses and lectures for school programs and academics), the digitized collection of books of the Sainte Geneviève library, the research documents of the ATILF laboratory (Analyse et Traitement Informatique de la langue Française – analysis and IT processing of the French Language), etc.

1.2 Missions

The boundaries of the preservation mandate are set by domestic laws:

- ✓ A statement (published on August, 7th 2006) which designates explicitly CINES as the national operator for the long term-preservation of electronic PhD theses;
- ✓ A mission letter (issued on February, 12th 2008) which reinforces the CINES mandate on digital preservation for four years.

In order to accomplish this official mission, CINES had to put a great number of resources together, with the objectives to:

- ✓ Create a dedicated department with a specific focus on access and preservation of digital objects on the long-term;
- ✓ Acquire and integrate specific skills (archivistic, project management, development competencies);
- ✓ Roll out a dedicated technical environment and share the infrastructure in place for the parallel computing activities;
- ✓ Be proactive and put in place an initiative to professionalize the activities and the business processes, improve the communication (conferences, trainings, etc.) and rationalize the strategy.

2. CERTIFICATION: GOAL AND STRATEGY

2.1 What is the rationale for certification?

Since the engagement letter issued by the Ministry of Higher Education and Research initially limits the mandate to a four years span, CINES must prove itself and lock the mission in the long term given the importance of the financial, technical and human resources required to execute it. A dedicated department has been set up for this purpose in 2008, with about ten engineers. CINES also put in place an important organization, which will only be relevant from an economic point of view if archived volumes increase significantly and CINES settles its legitimacy. Thus, the main objective of the approach is to get an official recognition that would allow to:

- ✓ Label the service;
- ✓ Legitimate its qualification;
- ✓ Become a professional in the French digital preservation community that cannot be ignored;

- ✓ Get a strong marketing point to develop the service with other communities;
- ✓ Communicate with the funding bodies.

One of most important criteria for certification is the viability over time of the mission entrusted to the organization. But in the Cines strategy, certification is a mean to legitimate its organization and establish the continuity of its mission, as well as a guarantee of fulfillment of the mission entrusted by the Ministry. These two conditions are obviously in conflict, and there are difficulties to change them into a virtuous circle.

In order to reach its certification goal, CINES bases its preservation and quality strategies on adaptation and use of standards such as:

- ✓ ISO 14 721 (Open Archival Information System);
- ✓ AFNOR NF Z42-013, French recommendations about conception and utilization of systems with data to preserve;
- ✓ Dublin Core (no qualified);
- ✓ A CINES standard based on ISAD-G and ISAAR (CPF) for project PDI;
- ✓ PAIMAS (Producer-Archive Interface Methodology Abstract Standard);
- ✓ Standard d'Echange de Données pour l'Archivage (SEDA), a French standard developed by DAF/DGME about archives exchanges (transaction and metadata schemes are described) [19];
- ✓ P2A - Politique et pratiques d'archivage – sphère publique, policy and pratices about preservation in a French public environment [13];
- ✓ Etc.

The certification process should be seen as an evaluation tool that encourages the preservation team to adopt more standards and to maintain a high quality service level.

2.2 The strategy toward certification

Much more than a simple management tool, the audit (ever more when internal) allows the repositories that adopt this technique to develop a deep knowledge of the way they operate, in a transverse manner.

In this context, CINES kicked off a certification process in which the main phases are:

- ✓ Permanent analysis and assessment of the different applicable standards to the CINES digital preservation department (started in 2008);
- ✓ Grant of the Data Seal of Approval accreditation (2008-2011);
- ✓ External audit (Ourouk consultants [22], Paris) for pre-certification, based on preservation standards : TRAC, DRAMBORA, ISO 16363 and ISO 14721 (2009);
- ✓ External audit for national agreement given by SIAF, a national service for coordination between Archives (2010);
- ✓ Participation in the EU funded APARSEN test audit project (Alliance for Permanent Access to the Records of Science Network) (2011);
- ✓ External audit for the CINES repository ISO 16363 certification (2012).

The timeline of the figure 1 shows this course of audits.

The strategy of CINES is to cover a large spectrum of standards and to increase the level of complexity required by the targeted certifications over time. Thus, the standards used for the first PAC certification were simple and based on auto-evaluations. The closer CINES gets to 2012 (the end of the span of the preservation mandate as per the mission letter issued by its Ministry), the more complex the certification standards are, to reflect the latest acquired experience and competencies.

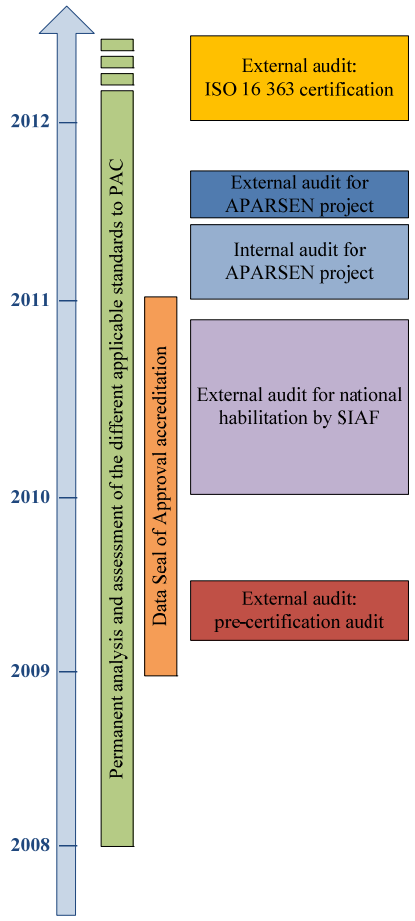


Figure 1. CINES strategy for certification.

Furthermore, the standard chosen for the "final certification" (the external audit in 2012) should sound familiar to the Ministry of Higher Education and Research. So, the future ISO 16 363 standard has been chosen for this: firstly, because it is an international standard and, secondly, because this draft describes

the preservation activity in CINES with more relevance than standard as ISO 9001:2008.

Even if the preservation mandate is renewed in 2012, the certification strategy should continue with periodic audits and improving quality assurance of the service.

3. QUALITY ASSURANCE: PREPARATION and IMPLEMENTATION

3.1 Policy

Any quality process needs policy and encouragement of the organization head. So, the quality strategy must be a part of the global organization strategy. In CINES, certification is a way to have a long term mission, one of the strategic goals.

In addition to these requirements, few elements were essential in the organization:

- ✓ First, communication is very important to avoid any rejection by the team. Consequently, the strategy for the evolution of the legal context (mission) was explained, regular meetings detailed the choices made, as well as plannings, relationships between team members and audit process, progress and results of audits, consequences of audits on daily work, etc. The active involvement of all the staff was decisive to identify nonconformities and execute an efficient and relevant audit.
- ✓ Second, transparency and honesty from the management are important too. At CINES, the certification approach is part of a constructive policy: its final objective is the realistic evaluation of the services provided to the communities, not a mean for the reorganization of the department. In other words, the independence and fairness of the auditors of the repository was a key factor in success.
- ✓ Last, the skills of the auditor are very important for the certification process to be fruitful and valuable. The knowledge and know-how of the auditors have been very much appreciated, during the internal and external audits.

3.2 The lack of relevant systems of reference – a difficulty for CINES

While producing a report, in 2008, on the state-of-the-art of existing certifications, CINES had highlighted the lack of specific, recognized business standards in the non-archivistic community. Year after year, a large growth of the certifications standards can be observed, among which:

- ✓ 2006-2007 : the methodology for self-assessment the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA), developed jointly by the Digital Curation Centre (DCC) and Digital Preservation Europe (DPE);
- ✓ 2007 : TRAC "Trustworthy Repositories Audit & Certification (TRAC) : criteria and checklist";
- ✓ 2009 : Data Seal of Approval, developed by DANS;
- ✓ 2011 : ISO 16 363 ("audit and certification of trustworthy digital repositories");
- ✓ 2011 : ISO 16 919 ("requirements for bodies providing audit and certification of candidate trustworthy digital repositories");
- ✓ 2011 – 2012 (?): French standard for certification based on the NF Z42-013 standard.

The early adoption of the criteria defined in audit systems of reference as well as other standards such as ISO 14 721 ("Reference Model for an Open Archival Information System") will help anticipating and resolving the problems bound to the development and production phases of digital repository infrastructures. By the mean of simple analysis, audits and/or self-evaluations, the regular study of preservation systems of reference can support the quality of the services provided.

3.3 Preliminaries: process documentation and DRAMBORA audit

Whatever the chosen standard, the documentation of the business processes is a prerequisite for any certification. From 2009 to 2010, CINES detailed its preservation activities through process maps and descriptive sheets. Fourteen processes have been identified and split into three categories: "management processes" (the processes that govern the operation of a system), "operational processes" (the processes that constitute the core business and create the primary value stream) and "supporting processes" (which support the core processes). The outcome of this initiative was partially presented during the iPRES2010 conference [1], and can be accessed online on the CINES website [2].

In the meantime, a first audit was executed internally in 2009, based on the DRAMBORA framework and online tool [9]. These works were coordinated by an archivist who had attended the specific training courses organized jointly by the Digital Curation Centre (DCC) and Digital Preservation Europe (DPE). The results of this audit have led the CINES to define a tailored risk management plan: thirty eight main risks have been identified and defined from the seventy eight risks listed in DRAMBORA. Each identified risks is assigned to a member of the digital preservation department. The risk management plan is reviewed twice a year by the whole staff, with the objective of reevaluating the probability and impact of each risk, defining action plans to mitigate them, etc.

4. CERTIFICATION: AUDIT TIME

Following the DRAMBORA audit and the completion of a substential part of the required documentation, CINES executed more audits based on more standards and a growing complexity.

4.1 2009: Pre-certification audit

When the first external audit of the CINES digital preservation repository was being negociated, no particular system of reference or standard had been imposed. The selected provider suggested to build a customized audit grid based on:

- ✓ The coming ISO 16 363 standard [8];
- ✓ TRAC [10];
- ✓ The checklist of the NESTOR project [15];

- ✓ The preservation policy audit grid as developed by the French Agence Nationale de la Sécurité des Systèmes d'Information [13];
- ✓ The OAIS conceptual model - ISO 14 721 [16];
- ✓ The French NF Z42-013 standard[14].

The resulting grid was filled by external consultants from evidences found in the documentation or observations from interviews with the staff. It was included in the final report which was structured as per the ISO 16 363 recommendations.

The workload associated to this audit represented nineteen man-days, and was done by two senior consultants. They interviewed the whole staff of the Digital Preservation Department as well as the members of CINES management.

The evidences provided to prove the compliance with the quality standard were:

- ✓ Documents (preservation policy, functional and technical specifications, process maps, event journals, etc.);
- ✓ Demonstrations of systems (functionalities like ingest, storage, data management, access, etc.);
- ✓ Documents and/or demonstrations of tools supporting business processes (ECM, etc.);
- ✓ Analysis of approaches for technology watching;
- ✓ Interviews.

From the report, actions plans have been defined and quickly put in place. The diagram below depicts the distribution of the criteria assesment for each type of recommendations.

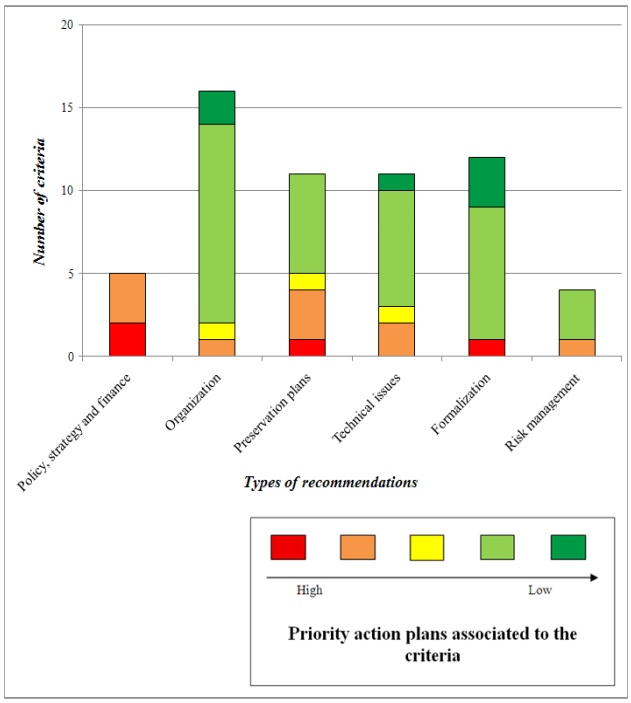


Figure 2. Criteria by types of recommendations (2009 external audit).

Thanks to this pre-certification audit, the CINES management and the middle management identified that priority actions were more related to policy, strategy, finance and preservation plans rather than on organizational aspects.

4.2 2010: Data Seal of Approval accreditation

The accreditation DSA Data Seal of Approval [5] is attributed to the digital preservation centers, for establishing quality assurance procedures to ensure accessibility and intelligibility of information entrusted to them.

The Data Seal of Approval was initially established by DANS in the Netherlands in 2007. A couple of years later, a number of institutions committed to durability in the preservation of research data took over to take the DSA to the european level. The current members of the DSA Board are:

- ✓ Alfred Wegener Institute (AWI), Germany;
- ✓ Data Archiving and Networked Services (DANS), Netherlands;
- ✓ UK Data Archive (UKDA), United Kingdom;
- ✓ Deutsche National Bibliotek (DNB), Germany;
- ✓ Max Plancke Institute (MPI), Netherlands/Germany;
- ✓ Inter-university Consortium for Political and Social Research, University of Michigan (ICPSR), United States;
- ✓ Centre Informatique National de l'Enseignement Supérieur (CINES), France.

By assigning the seal, they not only wish to guarantee the durability of the data concerned, but also to promote the goal of durable archiving in general.

It consists of sixteen guidelines split in three topics – data producer, repository and data users – with the objective to raise awareness on the importance of quality and ensure that in the future, research data can still be processed in a reliable manner, without entailing new thresholds, regulations or high costs.

To get the accreditation, which is based on trust, the repository has to submit a request on the web. In a description of the repository to be assessed, it should be explained that:

- ✓ The research data can be found on the Internet;
- ✓ The research data are accessible, while taking into account relevant legislation with regard to personal information and intellectual property of the data;
- ✓ The research data are available in a usable format;
- ✓ The research data are reliable;
- ✓ The research data can be referred to;
- ✓ The data producer is responsible for the quality of the digital research data;
- ✓ The data repository is responsible for the quality of storage and availability of the data: data management;
- ✓ The data consumer is responsible for the quality of use of the digital research data.

In 2009, CINES tested a first version of DSA with its digital preservation repository (PAC). It now complies with the 2010 guidelines version 1 set by the Data Seal of Approval Board. The

repository has therefore been granted the Data Seal of Approval for 2010 on March 15, 2011.

4.3 2010: external audit for national habilitation by SIAF

Since 2009, the French law allows organizations to store and preserve on the national territory some public records (non-heritage) provided that they have received an habilitation from SIAF (Service Interministériel des Archives de France). CINES, as a public institution and given the need expressed by its community, decided to position itself on this sector. The requirements from SIAF consist of twenty-two technical, operational, organisational, strategic and legal criteria. Such a level of demand relies on the standards of the domain such as ISO 14 721 and NF Z42-013.

In june 2010, CINES completed and sent a file to the Archives de France in order to officially request an habilitation. After few months of investigation, a group of eleven experts visited the CINES facilities and interviewed its representatives before issuing the habilitation on December 14th, 2010, for the next three years.

SIAF also provided a list of conditions and recommendations for the renewal of this habilitation, some of which had not be identified during the previous audits. CINES has already taken them into account in a specific action plan.

4.4 2011: internal audit for APARSEN project

APARSEN [6] is a european initiative led by the Alliance for Permanent Access to the Records of Science. Among the objectives of this EU funded project is the test audit of six digital repositories based on the ISO 16 363 standard, half of them being based in Europe, and the rest in the United States. This is also part of an initiative from the European Commission, started in 2010 to promote the rollout of a framework for the audit and certification of digital repositories. This framework would federate the different accreditation and certification project into three levels of recognition of the quality assurance effort done by institutions in charge of the preservation of the digital heritage, in increasing trustworthiness:

- ✓ Basic Certification through the Data Seal of Approval (DSA);
- ✓ Extended Certification through DSA plus additional publicly available self-audit with an external review based on ISO 16 363;
- ✓ Formal Certification after full external audit and certification based on ISO 16 363.

A memorandum of understanding [7] has been put together and signed by the different parties involved in this framework during the summer 2010.

The European datacentre being audited as part of the APARSEN projet were:

- ✓ The UK Data Archive (UKDA), United Kingdom;
- ✓ The Data Archiving and Networked Services (DANS), Netherlands;
- ✓ The Centre Informatique National de l'Enseignement Supérieur (CINES), France.

The experts in charge of the internal audit at CINES were :

- ✓ Olivier Rouchon, head of digital preservation department;
- ✓ Marion Massol, project manager (PAC);
- ✓ Jean-Pierre Théron, system administrator (PAC).

They were chosen because they have a good understanding and knowledge of the digital preservation process or the functional and technical management of preservation projects in PAC. Their recommendations in the final report have been made from assessment and observations. While trying to be as impartial as possible, the auditors have based their assessment on the following :

- ✓ Compliance in the 2009 external audit ;
- ✓ Improvement of compliance as part of the completed action and/or produced documents ;
- ✓ Gap between available documents and requested artefacts.

The internal audit performed as part of the APARSEN test audit project took place in four phases :

- ✓ A preliminary study (analysis of the reference document, definition of the scope of the audit, preparation of the main deliverable – report document in French, planning) ;
- ✓ An internal audit (evaluation and documentation of the criteria fulfillment in French, translation of the report in English language, additional interviews and verifications, gap analysis with the 2009 external audit report) ;
- ✓ The preparation of the documentation requested by the external APARSEN auditors ;
- ✓ The validation of the internal audit report/summary.

The workload for this internal audit was evaluated around sixty man-days.

The internal auditors set the functional scope of the audit on organisational and technical (management of digital objects, infrastructures, risk management in general, etc.) aspects.

The preliminary work in the internal audit anticipated a lack of evidences for the "access" functionalities as defined in the OAIS. The rationale for this is bound to the CINES policy/strategy to limit the access to archives to the sole data producers(aka transferring agencies), because most of them have their own websites for access and dissemination. The CINES repository will only provide a copy of their archives to the institutions in the event they have lost their copy or it has become obsolete. As of yet, there is no direct access to the archives for a larger community of users. A couple of studies have been conducted, and even if the technology is available in the CINES repository, there are no needs expressed by the user communities that would justify a complete process documentation and deployment.

The assesment of the criterias bound to security proved to be complex: in order to be relevant, such an evaluation must include the entire infrastructure used for digital preservation. Yet, a significant part of the infrastructure is shared with the HPC activities of the datacentre ; any security initiative has to include the whole CINES structure. Thus, such a work implies a lot of

efforts, resources involvements, etc. It has been started, under the resposibility of the RSSI (person Responsible of the Security of the Information System) but is not yet completed.

4.5 2011: external audit for APARSEN project

The external audit was executed on June 6th and 7th by twelve independant international experts nominated by the APARSEN consortium:

- ✓ Simon Lambert(United Kingdom);
- ✓ Donald Sawyer (USA, MD);
- ✓ Barbara Siermann (Holland);
- ✓ Robert Downs, CIESIN (USA, NY);
- ✓ David Giaretta(United Kingdom);
- ✓ Bruce Ambacher(USA, MD);
- ✓ John Garrett (USA);
- ✓ Terry Longstreth (USA, MD);
- ✓ Helen Tibbo (USA);
- ✓ Kevin Ashley (United Kingdom);
- ✓ Marie Waltz (USA, Chicago);
- ✓ Steve Hughes(USA, CA).

The audit started with an overview of the CINES approach and implementation to provide long term preservation of digital objects, followed by a visit of the facilities and a demonstration of the repository capabilities. Then, the auditors reviewed the report produced as part of the internal self-audit, and a question/answer session helped clarifying the remaining ambiguities.

As a conclusion, the auditors expressed remarks and recommendations for CINES to improve te quality of the services provided, where necessary.

The other objective of the APARSEN audit was to gather feedback from the institutions being audited as to the relevance and usability of the criteria listed in the standard. In some ways, it helped clarifying the ISO 16363 criteria evaluation system (methods/model for criteria appraisal, characterization of mandatory/optional compliances, etc.), as some questions were raised during the self-audit on this particular topic, and should be clarified in the final version of the standard to be published by the end of 2011.

The diagrams below (figures 3 and 4) show the progress made in the evaluation of the ISO 16363 criteria between 2009 and 2011 :

In the figures 3 and 4 above, the bubbles size, which are proportional to the labeled numbers, reflect the number of criteria with a given level of assesment and degree of importance, as per the respective evaluations. The area for improvement is clearly the criteria shown in red ; these have been adressed through action plans with high priorities. From the figure 2 (same legend for colors), we understand that the recommandations made for the criteria to be improved dealt with policy, strategy, finance, preservation plans and formalization of the activity.

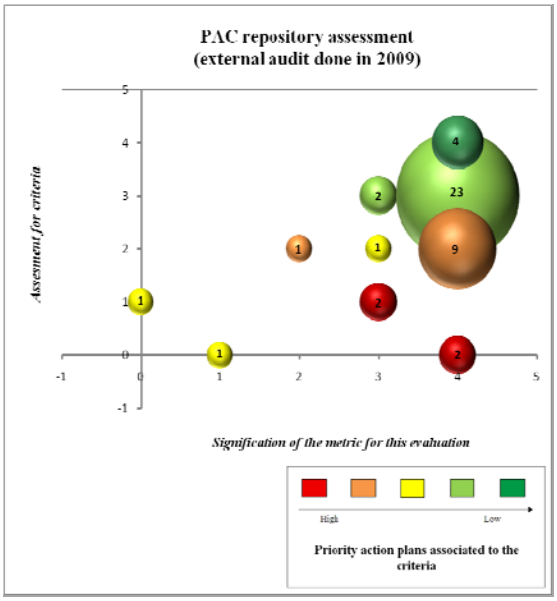


Figure 3. PAC repository assessment (2009 external audit).

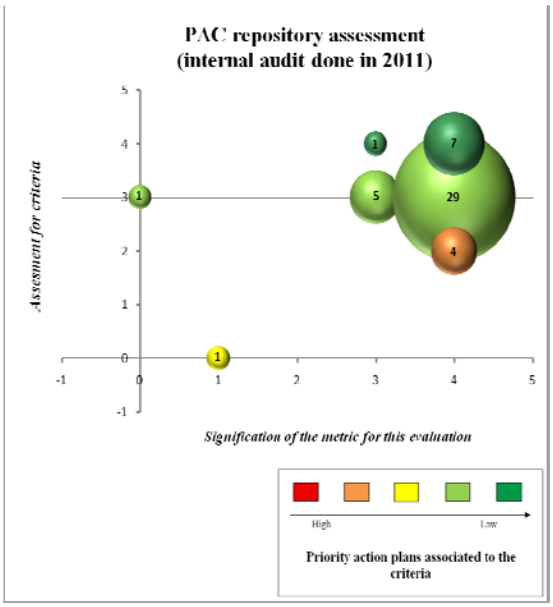


Figure 4. PAC repository assessment (2011 internal audit).

As a result, in 2009 and 2010, the CINES management focused on the improvement of these criteria. The outcome of the internal audit executed in 2011 confirms the efficiency of the action plans implemented over this period. Now, thanks to this last audit, CINES management can easily point that the criteria depicted in orange (assesment = 2 / 4) are the next to require to be actioned.

In the meantime, these audits allowed to find out some critical aspects which had never been met by the repository, among which:

- ✓ The gap in the level of knowledge within the team, and in the distribution of critical activities with the staff;

- ✓ The lack of end-to-end traceability for the object integrity during the ingest phase, that led to the obligation for the data producer to provide an initial checksum;
- ✓ The lack of formalization of some specific topics or processes (disaster recovery plan, business continuity plan, etc.).

As part of the deliverables, the auditors also provided additional reports that allowed to back the demand to ensure the continued existence of the CINES mandate and financing.

Even if the external auditors are not necessarily aware of the specific culture of the audited repository and staff, their fresh eyes on the project proved to be extremely valuable to argue the evidences, back some projects to improve quality (development of new internal modules for the repository, validation of contracts by a lawyer, etc.) and even suggest interesting things to look at and think about (potential strategic developments, internal communication improvement, etc.). From this point of view, the fact that the external auditors belonged to the digital preservation community and had a strong expertise of the domain was key to the success of the initiative.

5. CONCLUSION AND FUTURE WORK

The certification initiative that was kicked-off four years ago has been a great vector for the improvement of the quality of the services provided. Thanks to the documentation of the service activities, the problems bound to knowledge and competencies management between the members of the staff have been greatly resolved.

This experience and knowledge sharing goes beyond the sole PAC team and affects the whole community (shared technology watching, exchanges and feedback on issues, solutions, etc.). For this purpose, CINES participates in few workgroups, at the national level (groupe PIN [23], Commission Archivage Électronique de l'AAF [24], etc.) as well as the international level (Alliance for Permanent Access, Data Seal of Approval, EUDAT, etc.).

CINES is also willing to promote traceability and transparency toward its users : its preservation policy is available online on the CINES website, along with documentation intended for data producer to give an overview on the way archive projects are managed at CINES. This path through certifications contributes to reinforce the trust of data producers, funding bodies management or users toward the digital preservation platform and services.

Boosted by this experience, CINES is now willing to participate in standardazing activities, particularly in the certification domain. For this purpose, a member of the staff will join the ad hoc group responsible for the drafting of the yet to be AFNOR certification standard based on NF Z42-013 and NF Z40-350. Two other members of the staff are currently participating in the SEDA steering comittee, which objective is the improvement of the French standard d'échange de données pour l'archivage (SEDA) led by the Archives de France.

The outlook for 2012 and beyond relies on this outcome:

- ✓ Become a national reference in the digital preservation community;
- ✓ Get the ISO 16363 certification as soon as an organization provides audit and certification of candidate trustworthy digital repositories;
- ✓ Reinforce the participation in digital preservation standardization activities – at national and international levels.

In parallel to this certification approach, CINES is also moving its services toward the preservation of scientific data and datasets, which are produced by HPC systems for example. The CINES certification would indeed have a large impact of the success of such a project, which is planned to go live in 2012.

6. ACKNOWLEDGMENTS

The authors wish to thank:

- ✓ The DSA Board.
- ✓ The APARSEN project team and particularly the auditors, for their understanding and patience.

7. REFERENCES

[1] Marion Massol and Olivier Rouchon, *Quality insurance through business process management in a French Archive*. In the *International Conference on Preservation of Digital Objects iPRES2010* (Vienna, Austria, September, 19-24th, 2010).

[2] CINES web site : http://www.cines.fr/

[3] Section about long term preservation in CINES (CINES website): http://www.cines.fr/spip.php?rubrique219

[4] Charton, J. 2009. Comment... l'enseignement supérieur mutualise l'archivage. *01-Informatique - Business & Technologies* (May, 28th, 2009), 50-51. DOI=http://www.01net.com/editorial/503259/and-lenseignement-superieur-mutualise-larchivage/

[5] Data Seal of Approval : http://www.datasealofapproval.org/

[6] APARSEN website : http://www.alliancepermanentaccess.org/index.php/current-projects/aparsen/

[7] Memorandum of understanding (2010): http://www.trusteddigitalrepository.eu/Site/Trusted%20Digital%20Repository.html

[8] CCSDS 652.0-R-1, october 2009, *ISO 16 363: Audit and Certification of Trustworthy Digital Repositories*, draft recommended practice (red book) issue 1.

[9] DRAMBORA (Digital Repository Audit Method Based On Risk Assessment) : http://www.repositoryaudit.eu/

[10] RLG, February 2007, *Trustworthy Repositories Audit and Certification: Criteria and Checklist (TRAC)*, version 1.0, Chicago (USA).

[11] Simon Lambert and David Giaretta, *TRAC and ISO standardisation.* In the workshop on *Digital preservation of scientific information in a trusted environment* (Luxembourg, April, 28th, 2009).

[12] Robin L. Dale, *Certification of digital archives: an RLG update*. In the *DCC workshop on long-term curation within digital repositories* (Cambridge, USA, July, 6th, 2005).

[13] Prime minister's office, July 2006, *P2A – politique et pratiques d'archivage (sphère publique) et grille d'audit*, Paris (France) : http://www.ssi.gouv.fr/fr/bonnes-pratiques/outils-methodologiques/archivage-electronique-securise.html.

[14] M. Cathaly's commission (AFNOR Z40F), March 2009, *NF Z42-013*, Paris (France), Ed. Afnor.

[15] NESTOR (Network of Expertise in long-term STORage Working Group on Trusted Repositories Certification), 2006, *Catalogue of Criteria for Trusted Digital Long-term Repositories* : http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf

[16] CCSDS 650.0-B-1, January 2002, *ISO 14 721:2003, Reference Model for an Open Archival Information System (OAIS)*, blue book.

[17] *ISO 9001:2000 (X50-131)*, Editorial Afnor, France, 2000.

[18] *ISO 9001:2005*, Editorial Afnor, France, 2005.

[19] Ministère délégué au budget et à la réforme de l'Etat (direction Générale de la modernisation de l'Etat), Ministère de la culture et de la communication (direction des Archives de France), January 2010, *Standard d'échange de données pour l'archivage : transfert – communication – élimination – restitution*, Paris.

[20] ICA - International Council on Archives (Australasian Digital Recordkeeping Initiative), 2008, *principes et exigences fonctionnelles pour l'archivage dans un environnement électronique*. ISBN = 978-2-918004-00-4

[21] Direction générale de modernisation de l'Etat (DGME), *Référentiel général d'interopérabilité, version 1.0*, May 2009.

[22] Ourouk consultants : http://www.ourouk.fr

[23] PIN (Pérennisation de l'Information Numérique) workgroup web site : http://pin.association-aristote.fr/

[24] AAF (Association des Archivistes Français) web site : http://www.archivistes.org/

Users' Trust in Trusted Digital Repository Content

Devan Ray Donaldson
University of Michigan
School of Information
3339A North Quad, 105 S. State St.
Ann Arbor, MI 48109-1285
devand@umich.edu

ABSTRACT
Scholars who study trust in digital archives have largely focused their attention on the power of certification by third-party audit as a way to communicate trustworthiness to end-users. In doing so, they assume that the establishment of a network of trusted digital archives will create a climate of trust. But certification at the repository level also assumes the trustworthiness of digital objects within a repository; specifically that digital repository objects are authentic and reliable. This paper proposes the use of document-level seals of approval as a means of communicating to end-users about the trustworthiness of digital objects that is commensurate with specific user interaction. Implications of this proposed research stress the importance of assessing the 'real-world' impact of trust signals on users.

Categories and Subject Descriptors
H.1.2 [**Information Systems**]: User/Machine Systems – *human factors, human information processing.*

General Terms
Reliability, Experimentation, Human Factors, Verification.

Keywords
Authenticity, End-Users, Integrity, Trust, Trusted Digital Repositories.

1. INTRODUCTION
Archival scholars state that the trustworthiness (i.e., authenticity and reliability) of digital objects is important to users [5]. Criteria for repository certification include requirements for document level authenticity (i.e., integrity and identity) to ensure that users can be confident that they are interacting with authentic digital objects [14, 15]. Prior empirical research suggests that authenticity is important to end-users [4, 16]. Given that archival scholars, repository certification criteria, and prior empirical research all stress the importance of the trustworthiness of digital objects for end-users, it is surprising that research on how to communicate with end-users about archival trustworthiness is scant. End-users, *those not involved in the creation and preservation of the digital objects they use*, presumably know the least about the creation and maintenance of the digital objects they use, as compared to other classes of users such as creators or preservers. End-users have the greatest amount of uncertainty regarding whether or not a given digital object is authentic and reliable. Digital archivists must somehow provide end-users with information about their authenticity and reliability.

There are two potential ways to communicate with end-users about the trustworthiness of digital objects, specifically by: 1) exposing preservation metadata related to the authenticity and reliability of digital objects to end-users, or 2) using cues or symbols to denote the authenticity and reliability of digital objects for end-users. As a record, preservation metadata can be quite complex, sometimes providing more extensive data than the digital objects for which the preservation metadata were created. Given this, cues or symbols attesting to authenticity and reliability may be a more effective way of communicating to end-users about the trustworthiness of digital objects than exposing end-users to preservation metadata. This paper proposes seals of approval at the document level as one possible way to address this issue.

iPRES2011, Nov. 1–4, 2011, Singapore.

2. BACKGROUND
2.1 Archival Literature, Research and Users' Trust in Digital Objects
According to Duranti [5], archival trust involves two components: authenticity and reliability. Authenticity refers to the idea that a document is what it claims to be. Reliability refers to the idea that a record "can be treated as the fact of which it is evidence" [p. 7], and depends upon the form and procedure of creation for a record. Duranti wrote that both authenticity and reliability are important to users. Essentially, users need to know that a record [pp. 8-9]:

- is the same that was placed in the file by the creator of the file itself, and that it has been preserved in its integrity,
- is the same as the one that was transmitted to its addressee, and has not been manipulated or substituted in the course of the transmission,
- was made under controlled circumstances as part of the regular workflow,
- was made within a reasonable time after the occurrence of the facts it is about, and
- was generated by somebody who was competent to make that specific record, with either duty or the direct interest to make it accurate.

Empirical research on trust in digital objects has focused more on creators and preservers than end-users. MacNeil [9, p. 56] conducted case studies to ascertain which specific elements creators considered essential for verifying a record's authenticity. She also found out about the kinds of procedural controls exercised over systems and the records contained within them which, in the creators' view, support a presumption of authenticity. Donaldson and Conway [3] and Foscarini [7] found that preservers use preservation metadata to validate claims of authenticity for digital objects. Preservation metadata are "the information a repository uses to support the digital preservation process," and typically include some combination of descriptive, structural, technical and/or administrative metadata [12]. Little research has been done to assess whether or not preservation metadata could have trust value for end-users as they do for preservers in validating claims of authenticity for digital objects. This is important to consider because prior empirical research suggests that *end-users do have concerns about authenticity*. In Duff et al.'s [4] study, historians complained about copying errors, stating that such mistakes not only undermined belief in the continuing authenticity of a specific source, but also compromised the credibility of copies of other sources. Zhou [16] found that users of digitized archival materials were more likely to think those materials had been altered and were less confident in their own authenticity assessments than those who interacted with non-digital archival materials. If end-users have concerns about authenticity, how should archivists go about clarifying these concerns? How should archivists attest to the authenticity of the digital objects they preserve and make accessible for end-users? Should preservers provide end-users with preservation metadata because preservation metadata are what preservers use to validate document level authenticity claims? Or should preservers use symbols or cues such as seals of approval to denote the archival trustworthiness of digital objects?

2.2 Criteria for Repository Certification and Users' Trust in Digital Objects

In 2002, the RLG/OCLC Working Group on Digital Archive Attributes (WGDAA) [15] wrote the groundbreaking report entitled *Trusted Digital Repositories: Attributes and Responsibilities*. The working group defined a Trusted Digital Repository (TDR) as "one whose mission is to provide reliable, long-term access [of] managed digital resources to its designated community, now and in the future" [p. i]. The WGDAA also specified three levels of trust to apply to the establishment of TDRs, including [p. 9]: 1) How cultural institutions earn the trust of their designated communities, 2) How cultural institutions trust third-party providers, and *3) How users trust the documents provided to them by a repository*. Regarding the third identified trust level, the WGDAA wrote that users must be certain that a document received is the one requested and that a retrieved document can be verified to be the exact item deposited into the digital repository in the past. The working group recommended message authentication codes signed by trusted institutions and public key encryption systems as ways of addressing these concerns. While prior research suggests that preservers use checksums to establish the authenticity of digital objects [3, 7], research on the impact of such mechanisms on end-users' trust is limited in the literature.

Other closely-related means of establishing the trustworthiness of digital documents include certification of archives. The Archival Workshop Program Committee [1] characterized certification of archives as "[a] method by which an [a]rchive's customers could gain confidence in the authenticity, quality, and usefulness of digitally archived materials" [n. p.]. Subsequent certification standards endow a preservation repository with responsibility to ensure the authenticity of its digital objects through explicit criteria for repository level certification. For example, Trusted Repositories Audit and Certification (TRAC) [14] states in Section B6.10 that any repository that gains trusted status must enable the dissemination of authentic copies of the original or objects traceable to originals. TRAC explicitly states that, "[a] repository's users must be confident that they have an authentic copy of the original object, or that it is traceable in some auditable way to the original object" [p. 41]**.** Section A3.8 [p. 15] specifies that a repository must commit to defining, collecting, tracking, and providing, on demand, its information integrity measurements. Examples of mechanisms designed to address the integrity of digital documents include use of checksums at ingest and throughout the preservation process as well as keeping an explicit, complete, correct, and current record of the chain of custody for all digital content from the point of deposit forward (i.e., provenance). The criteria outlined in Sections A3.8 and B6.10 underscore the idea that part of repository level certification involves establishing the trustworthiness of digital documents, and establishing and maintaining trust in digital documents is accomplished using metadata. Given the importance of the association between repository level certification and document level authenticity and reliability outlined in standards for repository certification, more research needs to be done on how to effectively communicate with end-users about authenticity and reliability of digital objects.

The information needed to address Sections A3.8 and B6.10 of the TRAC criteria for repository certification would be best characterized as preservation metadata. Yet, as a record, preservation metadata can be quite extensive, sometimes more complex than the digital objects for which the preservation metadata were created. Cues or symbols attesting to authenticity and reliability such as seals of approval may be a more effective way of communicating to end-users about the trustworthiness of digital objects than exposing end-users to preservation metadata. Of course, seals of approval should only be granted to digital objects that have certain preservation metadata that can attest to their authenticity and reliability, even if those metadata are not exposed to end-users.

2.3 Research on the Effect of Repository Certification on Users

Little research has been conducted to understand the extent to which third-party audit and certification affect users' perceptions of trustworthiness. The CASPAR Consortium [2] conducted a study asking creators, curators and users of curated digital objects about the most important factors when determining whether to trust a repository. Among the most important factors, according to the study subjects, were: the track record of the repository's ability to curate objects; the repository's preservation of the audit trail for digital objects in its custody; and control of integrity within the repository. The findings are interesting because they indicate three important factors regarding users' trust in repositories that are interrelated and involve the authenticity and reliability of digital objects: how repositories curate digital objects, the metadata repositories collect for their digital objects, and control of integrity for digital objects.

2.4 Seals of Approval

While third-party certification checklists specify that TDRs be transparent in communicating audit results to the public, specific means of conveying information about the authenticity and reliability of digital objects is up to TDRs to decide. Research has shown that many users rely on cues and defer to heuristic rather than systematic processing when making trust judgments of digital objects found on the web [13]. As such, use of cues or signals to denote third-party certification may be an effective way to communicate this type of information and thereby build trust in digital objects with end-users.

Findings from empirical research in Human-Computer Interaction and E-Commerce support the idea that third-party seals of approval enhance users' trust. Fogg et al. [6] conducted a study with 2,500 participants and found that a website won credibility with users by showing seals of approval from known companies. Miyazaki and Krishnamurthy [11] conducted experiments designed to ascertain how online firm participation in Internet seal of approval programs affected consumers. They found that the presence of an Internet seal of approval logo resulted in higher levels of information disclosure and anticipated website patronage for consumers who experience relatively high levels of online shopping risk. Findings from these studies could be used to suggest the need for empirical research regarding the impact that document-level seals of approval could have on users' assessments of digital object trustworthiness.

Harmsen [8] describes a Data Seal of Approval program in which repositories complete an assessment document, undergo audit by a member of the international Data Seal of Approval Assessment Group, and publish the results of this assessment. Afterwards, repositories are allowed to use the logo of the data seal on their websites. To date, research on the Data Seal of Approval is very limited. Mitcham and Hardman [10] conducted a case study in which they outlined issues the Archaeology Data Service (ADS) faced in undertaking the repository certification process that precedes approved use of the seal. They also presented the potential benefits of Data Seal of Approval self-certification. One of the benefits of the Data Seal of Approval, the authors wrote, is enhancing the trust of their users. The effect of the Data Seal of Approval on ADS users was not examined in the case study. Since one of the perceived benefits of seals of approval is to positively influence end-users' trust in digital repositories, research ought to be done to examine the impact of seals of approval on end-users. Further, repository level certification says something specific about the trustworthiness of digital objects within a repository; specifically that digital objects are authentic and reliable. Document level seals of approval may be an appropriate way to communicate with end-users about the authenticity and reliability of digital objects.

3. RESEARCH DESIGN

To address the research question (How does a document-level seal of approval affect users' perceptions of trustworthiness of TDR content?), this paper proposes an exploratory experiment to investigate this phenomenon. The following proposed experiment focuses on digitized books as examples of TDR content.

3.1 Proposed Experiment

3.1.1 Hypothesis

Based upon prior research on seals of approval, this paper hypothesizes that participants will rate digitized books with seals of approval as more trustworthy than books without seals.

3.1.2 Design

This paper proposes use of an experimental design (see Table 1), selecting digitized books (B_n) that either have a seal of approval (denoted by the * symbol in Table 1) or do not. Participants will only see one version of each book. Book information content will be held constant for all conditions, ensuring that any effects would be due to the seals. All books used in this experiment will be randomly selected from a TDR. Seals will be assigned to books from the randomly-selected pool of TDR digitized books.

Treatment		Control
$*B_{1\text{-}10}B_{11\text{-}20}$ n=30	$B_{1\text{-}10}*B_{11\text{-}20}$ n=30	$B_{1\text{-}20}$ N=30

Table 1. Experimental design for assessing impact of document-level seals of approval on users' perceptions of trustworthiness of TDR content.

3.1.3 Participants and Procedure

Who to recruit for an experiment involving users of a TDR depends upon its designated community. Some designated communities are narrowly defined while others are loosely defined. Large-scale repositories that are not discipline-specific typically have very loosely-defined designated communities. This proposed experiment focuses on recruiting a sample of intended users of a TRAC-certified TDR - HathiTrust (HT) (http//:www.hathitrust.org). HathiTrust is based out of the University of Michigan but has over 50 institutional partners. The designated community for this TDR includes not only the students, faculty, and staff of all of its partners, but extends to include anyone with an Internet connection. A good place to start in terms of recruiting subjects for this proposed experiment would be undergraduate and graduate students at one of HT's partner institutions.

Each participant will be randomly assigned to a treatment or control group. To control for order effects, treatment and control groups will be subdivided. Thirty participants (n=30) will be recruited per subgroup to account for the law of large numbers. Participants will be asked to think about conducting a research task in which certain questions would need to be answered regarding eighteenth-century English literature. To simulate the seamless nature of cyberinfrastructure in which TDR content can be found, participants will be told to use the search engine provided to find books that could help them answer a series of questions. Participants will be able to type whatever search terms they choose, but every participant will be provided with the same set of search results (just in a different order). Half of the treatment group will see books with seals of approval added to their search result listing (odd-numbered) and the other half of the treatment group will see seals accompanying search result listings for even-numbered books. Each participant will assign a

trustworthiness rating (e.g., on a 5-point likert scale with 1 being not trustworthy at all and 5 being completely trustworthy) for each of the books they select.

4. EXPECTED OUTCOMES

Archival scholars, repository certification criteria, and prior empirical research suggest that end-users care about the archival trustworthiness of digital objects. So the question then becomes how to communicate with end-users about the trustworthiness of digital objects. This paper has argued for research to explore the impact of document-level seals of approval on users' perceptions of trustworthiness of TDR content. Empirical results that support the hypothesis that document-level seals of approval increase users' trust in digital objects would suggest that seals aid users in the way in which third-party certification was intended. Empirical results that fail to support this hypothesis would suggest that document-level seals of approval do not aid users in making trust judgments for digital objects and would need to be reexamined.

In an aggregated search environment, TDR content, which by definition has been upheld to best practices for authenticity, is listed alongside content in search results from other sources, which may or may not be upheld by the same standards. TDR administrators and designers need to develop effective ways of communicating with users about the trustworthiness of TDR content. This is a challenge, but if addressed, it could be of great benefit for users.

5. ACKNOWLEDGMENTS

I would like to acknowledge Kathleen Fear, Paul Conway, Paul Resnick, Eric Cook, Maciej Kos, Tracy Liu, Ann Zimmerman and the Archives Research Group at the School of Information for their comments and suggestions on previous drafts of this paper.

6. REFERENCES

[1] Archival workshop on ingest, identification, and certification standards. 1999. National Archives and Records Administration, http://nssdc.gsfc.nasa.gov/nost/isoas/awiics/ (accessed 10 August 2011).

[2] CASPAR Consortium. 2009. Report on Trusted Digital Repositories. Technical Report.

[3] Donaldson, D. R., and Conway, P. 2010. Implementing PREMIS: A Case Study of the Florida Digital Archive, *Library Hi Tech* 28(2): 273-289.

[4] Duff, W., Craig, B., and Cherry, J. 2004. Historians' Use of Archival Sources: Promises and Pitfalls of the Digital Age, *The Public Historian* 26(2): 7-22.

[5] Duranti, L. 1995. Authenticity and Reliability: The Concepts and their implications, *Archivaria* 39: 5-10.

[6] Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., and Tauber, E. R. 2003. How do users evaluate the credibility of web sites?: A study with over 2,500 participants. Paper presented at Proceedings of the 2003 conference on Designing for user experiences, San Francisco, California.

[7] Foscarini, F. 2008. "Cultures of Trust: Legal, Technical, and Archival Perspectives on the Use of Digital Signature Technologies," *Lecture Notes in Informatics (LNI),* vol. P-133: 37-47.

[8] Harmsen, H. 2008. Data seal of approval - assessment and review of the quality of operations for research data repositories. Paper presented at iPres, The British Library, http://www.bl.uk/ipres2008/presentations_day2/34_Harmsen.pdf (accessed 10 October 2010).

[9] MacNeil, H. 2000. "Providing Grounds for Trust: Developing Conceptual Requirements for the Long-term Preservation of Authentic Electronic Records," *Archivaria* 50: 52-78.

[10] Mitcham, J. and Hardman, C. 2010. ADS and the Data Seal of Approval – case study for the DCC, Digital Curation Centre, http://www.dcc.ac.uk/resources/case-studies/ads-dsa (accessed 29 September 2011).

[11] Miyazaki, A. D., and Krishnamurthy, S. 2002. Internet Seals of Approval: Effects on Online Privacy Policies and Consumer Perceptions. *Journal of Consumer Affairs* 36 (1): 28-49.

[12] PREMIS Editorial Committee. 2011. PREMIS data dictionary for preservation metadata *version 2.1.* Washington, DC: Library of Congress, http://www.loc.gov/standards/premis/v2/premis-2-1.pdf (accessed 30 August 2011).

[13] Rieh, S. Y. 2002. Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology* 53 (2): 145-61.

[14] 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 13 October 2010).

[15] RLG/OCLC Working Group on Digital Archive Attributes. 2002. *Trusted digital repositories: Attributes and responsibilities.* Mountain View, CA: RLG, http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf (accessed 13 October 2010).

[16] Zhou, X. 2005. A Comparison of Users' Response to Digital versus Physical Archival Material, Paper presented at the Society of American Archivists Annual Meeting, New Orleans, LA: 1-11.

Evaluation of a Large Migration Project

Alex Thirifays
Danish National Archives
Rigsdagsgården 9
1218 København K
+45 33 92 23 69
alt@ra.sa.dk

Anders Bo Nielsen
Danish National Archives
Rigsdagsgården 9
1218 København K
+45 33 92 83 26
abn@ra.sa.dk

Barbara Dokkedal
Danish National Archives
Rigsdagsgården 9
1218 København K
+45 33 95 46 89
bd@ra.sa.dk

ABSTRACT

The Danish National Archives (DNA) has ingested structurally heterogeneous public digital records since 1973. The year 2004 saw the creation of a new preservation standard into which it was decided to migrate the above mentioned archival holdings. The main objectives of this operation were to save data from technological obsolescence and to reduce the cost of both access and future migrations by streamlining the collection.

The project costs approximately 30 FSCs (one 'FSC'—Format and Structure Conversion—is the way the project's project management measured 1 person-year, and equals 1,291 person-hours). The total sum of purchasing software, hardware and external services amounted to around 135.000 Euros.

The project migrated data from both relational and hierarchical databases (for instance ERDMS and registries), and included the digitisation of audio, video as well as paper documentation. The registries counted for example the first Civil Registration System from 1968 and the State Tax Administration's final equation from 1970. Data and documentation made up a total of about 1.7 TB, consisted of 11,187 files scattered in almost 200 different structures, and constituted more than 2,000 information packages (IPs).

The overall technical objective of the migration was defined by the aforementioned preservation standard, which required:

- Common format for data files
- Common structure of documentation, metadata and documents
- Common format for documents (TIFF).

The project's main objectives were achieved, since all records were migrated, except the film collection. The goal of making access and future migrations easier was also reached, but a fully automatic migration of the collection is not yet entirely possible. The overall conclusion is that the migration project, which, to our knowledge, is the first of its kind, was of very high quality, both in terms of planning, execution and product.

A few main conclusions are:

- Standardisation of data is a prerequisite for an economically sound digital preservation: It took about 70 times longer to migrate an older, non-standardised IP, than a newer, standardised one.
- Inadequate feasibility studies remind us that the timing in digital preservation saves money: The condition of the magnetic tapes was examined through spot checking 10 years prior to the migration project and found satisfactory, which turned out to be an erroneous, expensive conclusion.
- The technical infrastructure suffered from a number of shortcomings and late decisions, resulting in precious loss of time.

Categories and Subject Descriptors

E.1.5 =E. Data, E.1 DATA STRUCTURES - Records

General Terms

Design, Documentation, Management, Measurement, Standardization, Verification.

1. INTRODUCTION

In the Performance Contract 2009-12 between The Danish National Archives and The Ministry of Culture it was agreed that the former write a self-evaluation of the migration[1] project, the Format and Structure Conversion Project (FSC), which was conducted by The Danish National Archives between 2005 and 2008.

The Format and Structure Conversion Project rendered the entire collection of digital records compliant with the Executive Order no. 342 of 11 March 2004 on information packages of preservation worthy data from IT-systems[2].

This Executive Order is now outdated and has been replaced by a new order[3].

A new "format and structure conversion" is hence expected. Therefore the experiences gained by the Danish National Archives from the FSC Project will serve as a valuable data basis for the planning of future migrations.

iPRES2011, Nov. 1–4, 2011, Singapore.

[1] The term "migration" is used instead of "conversion", cf. ISO 14721:2003, *Space data and information transfer systems – Open archival information system – Reference model*.

[2] In Denmark, (technical) Executive Orders constitute the submission standards.

[3] Annex 2 in the evaluation report – *Executive Oder no. 1007 of 20 August 2010 on information packages*.

This article aims to sum up the conclusions of the evaluation report[4], which is the first phase of a project[5] designed to develop strategy proposals for logical preservation and implementing concrete planning of future migrations within the framework of the Performance Contract 2009-12.

2. FSC PROJECT DESCRIPTION

2.1 Background

Since 1973, The Danish National Archives has received public digital records. Initially, The Danish National Archives only accepted electronic databases, but from the 1980s, it also accepted electronic filing systems[6]. At the end of the 1990s, it also became possible to receive electronic records and document management systems (ERDMS).

The older digital records existed in numerous formats and structures. This was because, until the end of the 1990s, the preservation standards available for digital archiving were not sufficiently defined. There were no format or structure descriptions available in a machine-readable format. As a result, it was difficult to test whether the digital records complied with the accompanied hardcopy format- and structure descriptions; this in turn led to the test not always being conducted, and when done, not always sufficient in extent.

Furthermore, until the end of the 1980s, there were no comprehensive requirements to the structure or format of the digital records submitted, nor were there any guidelines as to how this should be documented.

Hence, it was decided that all digital records should be migrated to the *preservation standard*, which coincided with the *submission standard*[7] applicable at that time. The standard describes which data structure and formats the records should be preserved in. It also describes what documentation and metadata should accompany each IP.

2.2 Objective and Introduction

The FSC project was a transformation[8] type migration project, with the overall objective of:

- Saving data from obsolete information-bearing structures and formats, and from decayed physical media (audio-visual records)
- Standardising data in such a way to make it feasible to enable future automated migrations and standardised accessibility to the records.

The FSC project took 4 calendar years to complete – from 2005-2008 – and included migration of all formats and structures that did not meet the requirements set forth in the aforementioned preservation standard. Briefly, this entailed that all hierarchical databases were migrated to relational ones; that all code pages were migrated to ISO 8859-1 (Latin 1); that all packed fields were unpacked; that all variable-length records were changed to non-variable; that all documentation was scanned and documented in accordance with the requirements, and that metadata were created.

The project cost just over 30 'FSC' person-years and the aggregate expenses for the purchase of software, hardware and external services amounted to around 135.000 Euros. The project had a steady staff head-count of between 10 and 15 employees. The project also had a steering committee and a reference group responsible for handling production and specialist questions. The steering group referred to The Danish National Archives' management.

2.3 The Records

The Danish National Archives' holdings[9] of non-standardised records, which were to be migrated, originated from three periods:

- **Period A:** submitted before the 1998 preservation standard[10] came into effect. These records were the most difficult and complicated to migrate because of their complex structures and unique formats.
- **Period B:** submitted in compliance with the 1998 preservation standard. Although these records were technically easier to migrate than those from period A, they were much bigger and included many code values and fields which had to be manually keyed in. Hence, the migration of this period's information packages demanded many resources as well.
- **Period C:** submitted in compliance with Circular No. 4 from 2000 and, after 2004, with Executive Order no. 342[11]. These records encompassed 942 information packages, but the number was constantly increasing as new submissions were received. The task of migrating that period's records was substantially easier than that of the preceding periods, since there was no large discrepancy between the 2 preservation standards, thus enabling a high degree of automated migration.

[4] http://www.sa.dk/media(3649,1030)/Final_report%2C_Evaluation_of_the_Format_and_Structure_Conversion_Project.doc

[5] Preservation Planning Project (PPP)

[6] "Electronic databases" is The Danish National Archives' term for (professional) databases. "Electronic filing systems" denotes databases with reference to hard-copy paper based case files.

[7] Annex 5 in the evaluation report – Executive Order no. *342 of 11 March 2004 on information packages worthy of preservation data from IT-systems*.

[8] Cf. OAIS terminology (Open Archival Information System (ISO 14721:2003, s. 1-13)): *A Digital Migration in which there is an alteration to the Content Information or Preservation Description Information of an Archival Information Package. For example, changing ASCII codes to UNICODE in a text document being preserved is a Transformation*.

[9] Note that the project was conducted by The Danish National Archives, and that the records were solely records from public administration and courts of law.

[10] Cf. *Electronic Archiving – The Danish National Archives system demands*, Danish National Archives, 1998, ISBN8774971778.

[11] These technical law texts represented quite similar submission standards, and, thus, preservation standards.

Table 1 - Overview of records in the FSC project

Record type	Submitted during the period	Data size (GB)	Number of files	Number of IPs
Period A[12]	1973-1998	171	3,109	650
Period B[13]	1998-2000	419	8,078	641
Period C[14]	2000-	1,187	-	942
Total	**-**	**1,777**	**11,187**	**2,233**

The records were categorized in information packages. An information package corresponds to a specific submission for a given period from an IT system, e.g. in the form of data from a registry for a 1-year period, or from an ERDMS for a 5-year period.

It has not been possible to accurately allocate the time used on each period separately, since some of the tasks performed were to the benefit of all three record types (e.g. transfer of preservation media). It has thus been necessary to make an artificial breakdown of the time spent. However, Table 2 shows a relatively accurate picture of the time spent.

Table 2 - Allocation of time used on each record type (A, B and C)

Record type	Hours	FSC-years	Time allocation (%)	Fraction of total collection (%)[15]
A	17,889	13.79	53	9.6
B	12,853	9.91	38	23.6
C	3,184	2.45	9	66.8
Total	**33,926**	**26.16**	**100**	**100**

The time spent here is not equivalent to all the time spent in the FSC project, since only tasks are included that can be directly related to the handling or processing of the three record types. Therefore, the time spent on migrating e.g. audio/visual records is not included in the breakdown.

2.4 The Technical Objectives

The main objective of the project was to transform data into information packages that complied with the requirements defined in the Executive Order no. 342; which, basically, means:

- **Common format for data files**
  - All tables must have fixed record length[16]
  - The content must be presented using a uniform code page: ISO 8859 Latin 1.
  - Fields in data files must be described using ISO standard data types such as NUM, REAL, STRING, DATE, etc.

- **Common Structure**
  - General information (documentation on the IT-systems' administrative function, structure and functionality)
  - Help tables (information on each submission, including context information, reference information and descriptive information)
  - Documents
  - Tables
  - Metadata, description of the information package's tables with field description and the mark-up of relations between them

- **Common format for documents** (TIFF)
  - Digitisation of paper documentation
  - Migration of digital documents (e.g. from Word to TIFF)

2.5 The Project

Most of the project required specialized knowledge, which only existed in the National Archives Preservation & Disposal Department; hence most of the tasks were performed internally. Some were outsourced, e.g. scanning of paper documentation for the digital records and the digitisation of analogue audio/visual records (sound and film).

The migration of the records to the preservation standard applicable at the time was done with a migration system which was developed in-house. Each migrated information package was transferred to preservation media in compliance with the implementation of The Danish National Archives' new media strategy of 2004[17]. All measures of the project's progression were performed in the archival database DAISY[18], while the registering of the records was performed partially in MARY[19] and partially in the DAISY.

In order to ensure demands for safety, security and confidentiality, the work was performed on an existing closed records network (Black Net) in the National Archives IT workshop.

This infrastructure formed the framework for the migration process, which can be simply illustrated as follows:

[12] Data files on tape and paper documentation.

[13] Data files on CDs and disks and paper documentation.

[14] ERDMS and registries.

[15] This column's numbers are based on each collection's data size in GB, cf. Table 1.

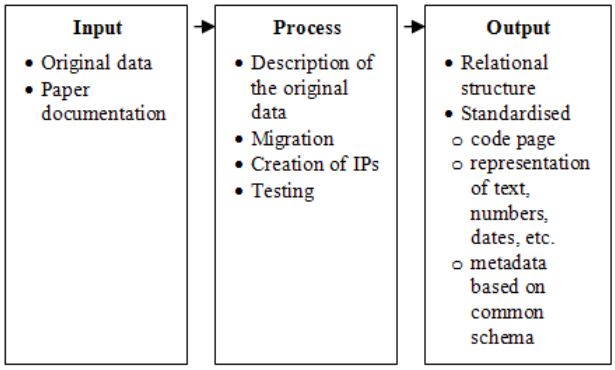
[16] According to the Executive Order no. 342, it is optional to choose between fixed and variable record lengths. However the FSC project systematically chose fixed record lengths.

[17] Cf. *New Implementations of the Media Strategy for The Danish National Archives, cf.* Jr. no. 2004-360-0005.

[18] DAISY is The Danish National Archives digital registry in which all user-oriented information about the creators of the IPs (authorities, companies, private individuals, etc.) as well as the IPs themselves is lodged. This database is the public's access to the holdings.

[19] MARY is The Danish National Archives preservation database used to monitor the condition of the digital collections. This database is for internal use only.

Figure 1 - FSCs migration process



The project went through a series of phases as described below.

1. Pre-project

The pre-project included the surveys that were prerequisite for estimating the budget of the actual project.

2. Restructuring the records to new preservation structure

The first task in the FSC project was aimed at transferring the older records into new preservation structures with a view to getting a unique identification of the information packages. The information packages were also transferred to preservation media.

3. Scanning

The task included the scanning preparation and the scanning of paper documentation (stored in ca. 400 archive boxes), which belonged to the old submissions. The paper documentation was the prerequisite for analysing and migrating the older records.

4. Registration

The task dealt with a basic and comprehensive registration and continuous updating and maintenance of data in the preservation database MARY used for monitoring the holdings. Moreover, the registration data in DAISY was made accessible in order to enable users to search the records.

5. Development of digital descriptions

The digital description included the development of a standard for the description of the records. This description standard was used to key-in the information from e.g. paper documentation on records descriptions. This was necessary for enabling the records to meet the preservation standard's requirements. The digital description standard was developed in XML with its own schema, and was used for achieving documentation on records in digital form. It was also used in order to get a digital form of the records' documentation. The digital form would enable the migration application to test and migrate data based on this description.

6. Development of migration application and ancillary modules

This task included the system development of the programmes that would support the migration process. The programmes included a central migration application (KonvOld), besides a number of modules intended to handle the variations in code pages, data structures, etc.

7. Further development of test, CD-burning and registration systems

The task included the further development of a number of already existing systems. These included programmes used for testing (TEA)[20], preservation media transfer (DEA)[21] and the export of automatically generated registration data to the preservation database, MARY.

8. Migration and test

This task was the central task in the project. The migration of all records was performed in this phase, which stretched over almost the entire project. It was by far the most resource demanding part of the project. The task was divided into the individual tasks, which records had to go through before finally being digitally described, migrated, tested, proofread, and placed on preservation media along with their corresponding digitised paper-based documentation.

An important part of this task was to prepare 2 documents establishing respectively the principles any migration would follow, and the routines any Convertor[22] would use. The migration principles and workflow descriptions should ensure that the records were uniform and consistently handled, so they were migrated in a standardized manner.

The migration was subdivided into three tasks:

a. Migration of unique records

This task included the migration of a variety of special records, which proved to be extremely difficult to handle and not readily processed by the central migration tool, KonvOld. The records also could not be handled by the digital description format. Hence there was a need for the development of record specific *ad hoc* tools and for a subsequent pre-migration, which prepared each unique record for the final migration using KonvOld. This required the use of experienced system developers with keen knowledge of the distinctive formats and structures. The unique records were prepared manually, one by one, in order to enable the migration to proceed automatically.

b. Migration of newer records

The task included the machine migration of a number of newer information packages supplied in accordance with the then recent preservation standard, Circular No. 4, and thus almost compliant with the preservation standard in force at the time (Executive Order no. 342).

c. Migration of analogue audio/visual records

Migration of analogue audio/visual records included the migration of The Danish National Archives' collection to the preservation standard applicable at the time.

[20] Test af Elektroniske Arkivalier - Test of Electronic Records

[21] Distribution af Elektroniske Arkivalier - Distribution of Electronic Records (CD burning and transfer programme)

[22] Convertors designates the staff used for migrating and testing.

3. Conclusions

3.1 Achieving the formal goal

3.1.1 Appropriation, budget and expenditure

The FSC project's appropriation was adhered to, but the project management had slightly under-budgeted the project (4.5 %):

Table 3 - Appropriation versus budget versus expenditure

	Appropriation	Budget	Expenditure
$	2,900,000	2,700,000	2,800,000

Consumption (hours)
The project planning had estimated using 39,053 hours or 30.11 man-years. The final number of hours used was 42,853 hours or 33.04 man-years. The difference is 3,800 hours, 2.9 man-years or just under 9 % which is partially compensated for in the monetary amount used (see above).

3.1.2 Records

Number
The project's main goal was achieved since all records were migrated, with the exception of the Danish National Archive's film collection, which had not been digitised. This was however accounted for in the revised project plan.

Quality and authenticity
The goal of making future migrations easier was achieved; but a fully automated migration of the collection is not possible: a high number of records – but not a large amount of data – had not been tested correctly and in consequence deviated from their original, appertaining preservation standard *before* the FSC project migrated them. This meant that some of those records also deviated from the preservation standard *after* migration. Maintaining the authenticity of the data instead of modifying them and stating how they were modified was a priority in accordance with the project management's preservation policy decisions. The partial lack of standardisation of those records is, hence, not due to poor migration, but to one of the Danish National Archive's preservation principles.

3.2 User evaluation of the migrated records

Concurrently with the migration project, an accessibility project developed an Access tool (SOFIA), which was launched in 2008. As of July 2010[23], this tool had been used to render 227 Information Packages 790 times.

No user feedback has been gathered in a systematic manner, neither regarding the access tool nor the accessed IPs. The feedback that has been received shows however that a number IPs are faulty, but according to the Access Department, the errors are easily corrected and modest in number.

The types of errors vary, but the most common ones are:

- Incorrectly converted data types, e.g. TIMESTAMP to STRING
- Inaccessible .TIFF context documentation documents
- Missing end date of the IP
- Occurrences of duplicates of document ID
- Missing validation of the IP after migration

[23] No newer numbers are available.

3.3 The Evaluator's Conclusions and Recommendations

After a thorough analysis of the FSC project, the Evaluator reached the following general conclusions taken from a broad review of the evaluated object. The conclusions are substantiated in more detail in the evaluation report.

The conclusions are of a summative nature and provide the basis for a review of:

- Future preservation policies and migrations
- Future projects in general

The following conclusions, which question a number of issues, do not dispute the overall conclusion, which is that the project was of very high standard regarding planning, execution and outcome.

3.3.1 Future preservation policy and migrations

Resource demanding projects

Almost 50 man-years is a ball park figure of how much time it took to complete the FSC project (circa 30 man-years) and the project's crucial prerequisites: the rescuing of magnetic tapes (circa 8 man-years) and the establishing of a preservation standard (circa 10 man-years). The figures should be interpreted with caution since not all assumptions of the calculation are fully taken into account.

The demand to stay within budget and the time constraint made it necessary to adjust quality

The FSC project budget was mostly adhered to. However, the budget was in fact an appropriation, and hence *should* be adhered to. There are several examples where the budget was met at the expense of other factors, such as product quality, performance, and quantity of migrated records.

- **The pre-project did not achieve all its stated objectives,** e.g. an in-depth analysis of records and the development of prototype programmes.
- **Digitisation of film, budgeted at DKK 700,000, was scrapped.** This decision was considered prudent by the project, as it was estimated that the analogue media the films were stored on were not at a risk preservation-wise. No documentation exists on the review of the media's shape or standard. The recommendation to place the media in optimal preservation conditions is not yet complied with.
- **Too few resources were allocated to purchase of hardware**. If the IT infrastructure had been in place on time, the project could have been completed several months earlier. Much frustration could have been avoided if e.g. the machine-processing time of the migration itself had been satisfactory.
- **Completing the handling of period C records was postponed** until after the end of the project.

- **There were vague plans to user-test the migrated information packages**, which was never done systematically. In the mean time, knowledge on the use of the records has been collected, and in general, The Danish National Archives' organizational entity, Access, has been very satisfied with the migrations. A log has been created including inappropriateness and errors that have been detected.
- **The Danish National Archives never managed to find the sufficient funds needed to get a satisfactory solution to the primary and foreign key errors,** which existed in some of the original databases and which had costly consequences for the accessibility of the migrated information packages.

Preservation policy principles can be expensive, but necessary to preserve our heritage

The Danish National Archives comply with the principle of never disposing of material if it already has been deemed worthy of preservation. Therefore, there is always the risk of using a lot of resources to preserve few damaged or faulty records at the expense of the majority of records. Examples:

- **In the recovery project**, which migrated damaged magnetic tapes, it cost on average 6.874 Euros to migrate just one magnetic tape, while the price for migrating a magnetic tape of good quality was 46 Euros (factor 149).
- **Unique records**. It took 70 times longer time to migrate older, non-standardised records (cf. period A) than newer, standardised ones (cf. period C). Hence, it took on average 0.23 hours to migrate *one* period C record, 13.72 hours for one for *one* period B record and 16.67 hours for *one* period A record. Measured per file, it took ca. 158 times longer to migrate the older records compared to the newer period C records (3.16 hours versus 0.02 hours). Per GB, it took 253.5 times longer to migrate one older record than a newer one (63.36 hours versus 0.25 hours).

Standardisation of data and tools is worthwhile

Standardisation of data is a prerequisite to ensure that digital preservation is economically sound, since the complexity and deviations are difficult to handle. The FSC project's main objective was to ensure that the entire collection of The Danish National Archives was standardised.

A complete standardisation would not only ensure the ease of developing for future accessibility and migration tools, but would also allow automated migrations.

- **Standardisation is expensive, but a good investment.** Standardisation of data is hence an investment, which can result in a modest amount of resources for migration and standardisation of the records. There are, however, still examples of records that cannot be processed in an automated fashion by migration tools.
- **System development of tools for non-standardised records is expensive.**
    - Development of tools needed to handle the older records (period A and B), accounted for over 80 % of the total system development costs.
    - 175 programmes were developed to handle the 167 records with unique structures and formats.
- **The flexibility of the central migration application (KonvOld) allowed for easier addition of modules and thereby ease of handling many formats.** Only 167 out of 1,291 records (just under 13 %) were to be pre-migrated by *ad hoc* tools before they were suitable as data input to KonvOld. The remainder of the data could be used directly as standard input. KonvOld handled thus the automated migration of ca. 87 % of records. This was only possible due to the ongoing, iterative optimisation and expansion of KonvOld.
- **A standardised, system-independent format was chosen for preparing the digital descriptions:** An XML schema. This made it possible to use alternatives to InfoPath tools used for reading and maintaining the digital descriptions.
- **Two preservation databases are one too many.** Having to do (double) registering in MARY and DAISY required much manual resources - it took ca. 1 person-year. It is, however, important to note that it might not have been economically feasible to merge the two systems within the scope of the FSC project.

Software development method depends on knowledge of data

- **When data are not known, it is best to use iterative development.** The migration application, KonvOld, was developed iteratively. The main reason for this was that the records to be handled were of very different structure and format and were not known beforehand. The iterative development ensured the possibility of automated processing of 87 % of the records, which was a huge advantage to the project.
- **When we assume we know the data, the tool should be fully developed before it is used.** The test application, TEA, too underwent iterative development and existed in several consecutive versions, which to varying degrees met the requirements associated with the applicable preservation standard. This had an adverse consequence: the records were received, tested and approved (with errors) by various versions of the same tool, depending on when the submission actually occurred. Cf. the points below. It is unknown whether it is possible to avoid versioning of test tools.

Quality control is expensive but should be appropriate in extent and quality

- **It is costly to perform many random checks, but risky (and potentially more costly) to perform too few.** Random checks demand many resources but are necessary to perform on a large scale. There have been too many instances of rash conclusions – cf. the bullet below, '**Feasibility studies'**, pertaining to an insufficient random check of the magnetic tapes.
- **Test tools must test exactly what is described in the preservation standard,** otherwise, the records are not tested in compliance with the standard, which makes it impossible to perform future automated migrations and accessibility. As a minimum, it is required to know

which version of the test tool the individual information package is tested with, and this version must be well documented in order to describe to what degree it lives up to the preservation standard.

Performance – software and hardware should efficiently be able to process a given amount of data

The technical infrastructure suffered from a number of shortcomings and late decisions that cost many resources.

- **Hardware lacked power.** The migration of records could have been performed much faster. Although a late decision on the purchase of new machinery helped to meet the deadline, it was still not sufficient to make up for lost time.
- **Tools performed poorly in certain contexts.**
  - When using InfoPath for the digital descriptions, the speed of loading and editing was very slow. This was especially the case for large records containing many tables, fields and long code lists.
  - KonvOld had the same problem of controlling the migration of large data quantities. The reason was that KonvOld did not use a database when controlling output data.

Quality – the better the input data, the cheaper the migration

It cost a lot of resources to improve on the poor quality of the records prior to migration, which emphasises the need for keen supervision and testing when receiving records.

- **1.6 % of the magnetic tapes cost 75 % of the total project amount** in the recovery project, because the tapes were in so poor shape that only an external specialist was able to restore the data.
- **The quality of the input data was poor**, and in some instances irreparable, which to a certain extent is reflected by the output data, as not all of these live up to their preservation standard. This meant that for some information packages, there was a need for exemptions from the rules, otherwise automatic processing is impossible.
- **The quality of the documentation was not satisfactory** and did not always give a complete description of the records. This meant that many resources were used to examine the formats, structures and content of the records in order to optimise the documentation in its digital form.

3.3.2 Future Projects in General

Project planning – Tight control and loose methods

The FSC project was well planned. Nevertheless, there was always a clear understanding of the need for an ongoing learning curve. This enabled an improved decision making basis, but meant that the project ran the risk of deviating from the chartered course. This, however, did not occur.

The knowledge policy was efficient

The FSC project created a favourable environment for knowledge sharing; however, some of this knowledge was lost.

- **Migration principles and work-flow descriptions.** Since the records were very different in type, it was necessary to accumulate and share a great deal of knowledge. This knowledge was preserved through meticulous and continuous written documentation of the working procedures. The documents were discussed and updated at the weekly project meetings.
- **Quick access to information.** Several factors were of great value to the daily work within the project, namely: online access to the majority of the project's information, the 2 documents mentioned above, interactive access to the records' scanned documents via MARY, and the digital description.
- **Temporary employment and loss of specialized knowledge.** Many of the project staff members gained both IT professional and professional archival expertise on digital preservation. This knowledge was largely lost when the employment contracts expired, but *qua* the standardization of the records, that knowledge is no longer needed.
- **Vulnerability of the project.** Even though this did not give rise to major problems, and the vulnerabilities were partially addressed in the project planning risk assessment, the project remained vulnerable in two areas:
  - **Terms of employment**. The project risked seeing the temporary employees leave prematurely when nearing the end of the project, which would result in loss of knowledge at a critical time.
  - **Staffing**. The project operated with a skeleton staff. However, there was no contingency for illness or vacations, etc.

Focus on methodology and daily work

The documents describing migration principles and work-flows were invaluable in the daily work. They ensured a uniform, high quality and efficient processing of the records.

- **It was not possible to determine the migration methodology beforehand,** since the feasibility studies did not sufficiently document the nature of the records. The three documents, which constituted the migration methodology, were hence "dynamic" documents that reflected a process in constant change.
- **This created an explicit need for knowledge sharing,** which was formalised at the weekly meetings and in the collaboration forums in which the convertors worked together two at a time. All this led to updated, user-friendly and accessible documentation of methodology.

Sustained precision

The large amount of routine work and resource demanding manual tasks (over 6,000 hours) constituted a bit of a challenge, with much that could have gone wrong and time that could have been wasted. Based on the project staff feedback and on the machine-controlled data, we can ascertain that the tasks were handled to a very high level of sustained precision. The reason for this can be found in:

- **Control**. It was possible to machine control the production of some tasks. In addition, the convertors worked together in pairs.
- **Professional pride**. Interview with the temporary project staff showed that this pride was decisive in maintaining the high quality of the manual work performed.
- **Delegation of responsibility** by the aid of *serial* allocation. The FSC project allocated to the Convertors their "own" series of records, for which they were responsible. These series were semantically similar, enabling recognition and identical treatment, and the sharing of responsibility partially explains the diligence with which the work was performed.
- **Passion and enthusiasm**. The organisation managed to create and maintain a high level of motivation amongst the temporary staff. Whether this was due to the project manager, the project staff, their common situation (temporary project employment) or a fourth reason, is difficult to assess. However, it was said that there was a very good chemistry amongst the staff and a good social environment.
- **The staffing committee succeeded** in hiring IT professionals who had earlier worked with projects that required the same level of structured work as in this project.

Insufficient focus on the FSC project's ancillary activities

In general, there was too little focus on the activities that improved on, documented and quality assured the project. One example illustrates this:

- **Feasibility studies**. **Several instances of insufficient feasibility studies which, if done, could have provided potential savings and efficiencies.**
    - **The condition of the magnetic tapes had already been examined in 1995 through a random check. The tapes' condition was found to be suitable, which it was not**. Therefore, a migration of the media's data was not initiated at that time, which in turn resulted in the huge expenses incurred to salvage the tapes 10 years later (cf. on average 6.874 Euros to migrate just one damaged magnetic tape).
    - **The pre-project did not have enough time to examine a suitable sample of the data to be migrated**, which meant that it was not possible to produce a detailed requirements specification for KonvOld, which had to be developed simultaneously with the migration process (iteration). The fact that this turned out to be a very advantageous solution could not have been predicted.
    - **Preparation for the system development was perhaps inadequate.** If there had been set aside more time for surveying the market and trying out various commercial or open source tools, it might have been possible for the FSC project to save resources on system development.

Lengthy decision-making processes

- **In certain instances, the chain of command was bureaucratic and impeded the quick implementation of decisions.**
    - **Change management**. Internal improvements in work processes gave rise to a greater need for making backups than originally planned. In spite of arduous meetings and negotiations, the FSC project did not succeed in getting more resources for backup and hence the project resorted to doing the backups of the production data on loose media.
    - **Purchase of additional machines**. A need arose for more machine power than originally anticipated. The decision on which machines to buy took a long time, which delayed the completion of the project.
    - **Optimal configuration of workstations** (images) was never done by The Danish National Archives' operations department and was left to be completed by the FSC staff.
- **Fundamental decisions regarding migration methods (waivers and deviations from preservation standard) had an approximately 14-day turnaround time, which is both acceptable and necessary.**

Outsourcing

- **Outsourcing of tasks does not entail freeing up of resources to internal activities pertaining to the task.** The major share of the expenses for scanning of documentation, which was outsourced, was taken from sub-tasks performed in-house (55 %).
- **Alternatives to in-house development.** The extent of outsourcing tool development should be considered. The obvious advantage of outsourcing is profitability, while the disadvantage is loss of control, difficulty of integration in the preservation environment, and efforts made in order to find a suitable vendor for such highly specialised tools.

Insufficient communication among the various projects

Despite a common steering group for the FSC project and the concurrent Accessibility project (TGP), the lack of communication had costly consequences.

- Migration of period A and B records allowed (after approved exemptions) the production of information packages, which included errors in relations, i.e. duplicates in primary keys and lack of foreign keys. It was decided not to use a machine-based, technical correction of the errors, e.g. by designating "dummy" key values. That solution would be detrimental to the data's authenticity, but would have curtailed the development time of the accessibility tool, SOFIA, by around 1 year.
- During the process of buying hardware, the IT department had a fall out with the FSC project staff. A better cooperation would have ensured a more powerful hardware performance than what was used earlier in the project, which would have ensured faster migration.

- There was inadequate integration between the components for preservation (FSC), registration (Access department with DAISY) and accessibility (TGP). It was e.g. not possible to automatically import data from KonvOld to DAISY.

Perhaps the lack of communication can be explained by the fact that the project was not properly aligned with the rest of the organisation. Another reason could be that there was not enough time allocated to meetings amongst the various projects' members. In the long run, it is recommended to completely integrate the preservation environment, as is directed by e.g. the OAIS model in which digital preservation – from data collection to logical preservation to accessibility – is considered as one organisational unit.

4. REFERENCES

[1] ANSI/ARMA, 2007, *The Digital Records Conversion Process: Program Planning, requirements, Procedures*, ARMA International, 13725 West 109th Street, Suite 101, Lenexa, KS 66215, 913.341.3808.

[2] Consultative Committee for Space Data Systems, 2002, *Reference Model for an Open Archival Information System (OAIS)*, CCSDS Secretariat, Space Communications and Navigation Office, 7L70, Space Operations Mission Directorate, NASA Headquarters, Washington, DC 20546-0001, USA Sannella, M. J. 1994. *Constraint Satisfaction and Debugging for Interactive User Interfaces*. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington. DOI = http://public.ccsds.org/publications/archive/650x0b1.PDF

[3] Consultative Committee for Space Data Systems, 2004, *Producer-Archive Interface Methodology Abstract Standard*, CCSDS Secretariat, Space Communications and Navigation Office, 7L70, Space Operations Mission Directorate, NASA Headquarters, Washington, DC 20546-0001, USA Sannella, M. J. 1994. *Constraint Satisfaction and Debugging for Interactive User Interfaces*. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington. DOI = http://public.ccsds.org/publications/archive/651x0m1.pdf

[4] CRL, The Center for Research Libraries, OCLC Online Computer Library Center, Inc., 2007. *Trustworthy Repositories Audit & Certification: Criteria and Checlist*. 6050 South Kenwood Avenue, Chicago, Illinois, 60637-2804 USA

[5] DNA, The Danish National Archives, 1998, *Electronic Archiving – The Danish National Archives system demands*, ISBN8774971778.

[6] Dollar, Charles M., 2000. *Autrhentic Electronic Records: Strategies for Long-Term Access*. Cohasset Associates, Inc. Chicago, Illinois, USA

[7] ISO 15489-1:2001, *Information and documentation -- Records management -- Part 1: General*

[8] ISO 23081-1:2006, *Information and documentation -- Records management processes -- Metadata for records -- Part 1: Principles*

[9] ISO/DIS 13008, 2011, *Information and documentation -- Digital records conversion and migration process*

[10] ISO/IEC 29121:2009, *Information technology -- Digitally recorded media for information interchange and storage -- Data migration method for DVD-R, DVD-RW, DVD-RAM, +R, and +RW disks*

Developing a Robust Migration Workflow for Preserving and Curating Hand-held Media

Angela Dappert
Digital Preservation Coalition
Innovation Centre, York University
Science Park, Heslington,
York YO10 5DG
angela@dpconline.org

Andrew Jackson
The British Library
Boston Spa, Wetherby
West Yorkshire, LS23 7BQ, UK
+44 (0) 1937 546602
Andrew.Jackson@bl.uk

Akiko Kimura
The British Library
96 Euston Road
London, NW1 2DB, UK
+44 (0) 20 7412 7214
Akiko.Kimura@bl.uk

ABSTRACT
Many memory institutions hold large collections of hand-held media, which can comprise hundreds of terabytes of data spread over many thousands of data-carriers. Many of these carriers are at risk of significant physical degradation over time, depending on their composition. Unfortunately, handling them manually is enormously time consuming and so a full and frequent evaluation of their condition is extremely expensive. It is, therefore, important to develop scalable processes for stabilizing them onto backed-up online storage where they can be subject to high-quality digital preservation management. This goes hand in hand with the need to establish efficient, standardized ways of recording metadata and to deal with defective data-carriers. This paper discusses processing approaches, workflows, technical set-up, software solutions and touches on staffing needs for the stabilization process. We have experimented with different disk copying robots, defined our metadata, and addressed storage issues to scale stabilization to the vast quantities of digital objects on hand-held data-carriers that need to be preserved. Working closely with the content curators, we have been able to build a robust data migration workflow and have stabilized over 16 terabytes of data in a scalable and economical manner.

Categories and Subject Descriptors
H.3.2 [**Information Storage**]; H.3.6 [**Library Automation**]; H.3.7 [**Digital Libraries**]; I.7 [**DOCUMENT AND TEXT PROCESSING**]; J.7 [**COMPUTERS IN OTHER SYSTEMS**]

General Terms
Management, Documentation, Performance, Design, Experimentation

Keywords
Data-carrier stabilization, disk-copying robot, digital preservation, auto loader

iPRES2011, Nov. 1–4, 2011, Singapore.

1 INTRODUCTION
Digital objects typically undergo several processing steps before they can be considered well-managed. For practical purposes, we use 4 coarse stages to describe how well managed a digital object or collection is. This is a simplification of the criteria one might use in a full risk assessment. For example, it conflates several criteria and the states don't necessarily develop in exactly this order. Nevertheless, it gives a useful pragmatic classification to assess the preservation quality of a set of collections.

The lowest-rated category is that of handheld data-carriers and is considered absolutely unsatisfactory for digital preservation purposes. Hand-held data-carriers tend to decay quickly (particularly writable or re-writable carriers that use optical dyes) and their rendering devices become obsolete relatively quickly. It is not always feasible to properly quality assess them upon arrival, due to the high manual overhead associated with checking each item. This also means that there is often only a single physical copy held, with no backup in case it should get damaged. Online storage is considered more resilient, easier to check for deterioration, easier to refresh if deterioration is detected and easier to manage remotely. For all these reasons it is very desirable to move digital objects as quickly as possible into the second category, the bit-stable category.

Table 1. 4-category digital object status progression

Unsatisfactory object status	**Bit-stable object status**	**Content stable object status**	**Archival object status**
Hand-held carriers	Content has been transferred onto managed hard disk storage. Storage is backed up. Checksums have been calculated.	Content has been QA'ed. Metadata has been produced and QA'ed. File formats have been identified. Representation Information has been deposited.	Automatic check for corruption via checksums. Automatic replication over remote locations. Digital signatures. Integration with the catalogue.
	Step 1	Step 2	Step 3

To do this, the digital objects must be transferred onto managed hard disk storage and storage that is backed up. We refer to this migration process as data-carrier stabilization. In addition to preservation concerns there are other functional requirements to hold digital objects on managed hard disk storage: to render the content searchable, to provide remote access, to replicate content in several locations, etc. The next two steps of creating content stable and archival object status, as described in Table 1, are just as important. But this paper will focus on issues involved in this first step of creating bit-stable objects.

The British Library has numerous collections, which exist to a significant degree on hand-held data-carriers. They vary in their characteristics, their preservation needs, and the eventual usage of their content. For example, in the Endangered Archives Programme [27], funded by Arcadia [4], valuable digitized material is sent to the British Library from projects all over the world. Due to local variability in technology, the content and the data-carriers on which the content is transmitted vary greatly; they are not always satisfactorily QA'ed *before* they are sent in, file formats vary, and they don't arrive uniformly catalogued. This means that content and metadata often need to be edited and restructured by curators to create a consistent collection, and the stabilized material needs to be easily findable, accessible by curators, and needs to be associated with metadata that may be held offline. While the data-carriers themselves are not valuable artifacts that need to be preserved, their content is and the distribution of the content over several data-carriers is sometimes associated with important semantic distinctions. In contrast, most of the CDs and DVDs in the Sound & Vision collection are published, mass-replicated disks, and have a much higher life expectancy than (re)writable disks. Furthermore, the disks themselves are valuable artifacts that need to be preserved and, frequently, a data-carrier corresponds to a uniformly catalogued work. Personal Digital Archives of persons of interest to the heritage community often contain manually created hand-held data-carriers. Like the other non-commercial data-carriers, they can vary greatly in structure and quality, but also form valuable artifacts in and of themselves. They typically don't require mass stabilization but may be analyzed in detail for nuances, even for deleted files. These differences result in different approaches to stabilizing their content.

In this paper we explore the experiences gained during our work on the Endangered Archives Programme (EAP), with a particular focus on optical disk processing (rather than tapes, external hard-drives, etc.). All together, the EAP collection currently contains 67 terabytes of data, on approximately 18,000 optical data-carriers, with more to arrive for a further 8 years. In the future, a larger proportion will arrive on more efficient data-carriers, such as external hard-drives. But this was not always practical in the past as early external hard-drives were not found to be robust enough to survive transport and indefinite storage. Much of the material has therefore been submitted via writable CD-ROM and DVD disks, and manually handling these data-carriers has proven to be enormously time consuming. It is, therefore, important to develop scalable processes, and to establish efficient, standardized ways of recording metadata and to deal with defective data-carriers.

The goals of the Endangered Archives Stabilization Project were:

- to move the digital material from hand-held carriers onto backed-up online storage in order to stabilize it,
- to develop workflows so that future accessions can be immediately made available on online storage,
- to determine the best technical set-up, processing approach and software solutions,
- to determine staffing needs for the stabilization process.

We have experimented with different disk-copying robots, developed stabilization processes, defined metadata, and addressed storage issues to scale stabilization to the large quantities of digital objects on hand-held data-carriers that need to be preserved.

2 MIGRATION STRATEGY

In general, we wish to maintain the authenticity of the original item as closely as possible. Ideally, therefore, we would aim to perform a reversible migration, such that the digital entity we create from the original data-carrier could be used to create a new data-carrier that is functionally equivalent to the original.

To understand how this might best be achieved, we first summarise how data-carrying media are designed. In order to function, any data-carrier capable of carrying more than one bitstream must use some kind of container format to arrange the bitstreams on the carrier in a way which allows them to be reliably disentangled when read. This is achieved in a range of different ways depending on the media, e.g. by using disk partition maps and file-systems. Necessarily, in order to allow the bitstreams to be distinguished, these container formats must also specify some metadata such as the filename associated with the bitstream, where on the disk it can be found, and how big it is. Usually, the container metadata also includes checksums and error-correction codes to help compensate for any bit loss during creation, aging or use of the media.

By definition, it is impossible to extract the individual bitstreams from the carrier without also stripping away the container. If we are fully confident that we are aware of all of the potential metadata that we might wish to keep, then this information can be extracted along with the bitstreams. But evaluating the auxiliary container-level metadata is time consuming, and if we are forced to make this evaluation directly from the physical media then the media handling process becomes extremely difficult to scale.

Fortunately, this bottleneck can usually be avoided by creating disk images. Here, rather than extracting bitstreams directly to files in a new file system, we attempt to extract a single large bitstream that represents a precise copy of both the contained bitstreams and the container. In this way, we preserve the logical content as completely and as closely to the original as possible. Note this does not necessarily preserve the precise physical layout of the data. For example, a hard disk cloned in this manner will contain the same information as the original, but will not have the same degree of data fragmentation, as the block-level data layout will have changed. However, the clone is logically equivalent to the source disk, and this is entirely sufficient for our purposes.

Critically, this approach also allows us to proceed quickly, migrating the content as soon as possible while minimizing the risk of discarding any data or metadata during bitstream extraction or container transformation. By creating a disk image we can move the original submissions onto safer storage without compromising the authenticity of the originals. This approach is also common in the digital forensics area, and well-established practices are in place for many types of media [7].

2.1 Variations in Carrier Type

While the broad strategy of making disk images is a sound one, there are a number of practical difficulties implementing this ap-

proach due to the variations in the types of disk and the degree to which the disks conform to the appropriate standards.

The variation in disk formats arises due to the complex history of the medium, and the ways in which the form has been extended or modified to cover different use cases. The original Red Book [12] standard from Phillips specified how to construct a digital audio compact disk, with raw audio bitstreams arranged into a series of session and tracks, along with the physical layout and analogue tolerances to which this format should be constructed. In the following years, a wide range of other standards were published (the so-called Rainbow Books [30]), covering extensions and modifications to this base format, such as CD-ROMs for data, mixed-mode audio and data disks, extended embedded metadata, technical protection mechanisms, and so on.

Since then, and in reaction to the complexity of this group of standards, the vendor community has worked to standardize the way in which the data is laid out upon the disk, via the Universal Disk Format [16]. Both DVD and Blue Ray media use this disk format, which specifies just one container format, but captures the different media use cases in the standardization of the bitstreams within the container, rather than via the structure of the container itself. This is not done for reasons of preservation, but for reasons of ease of creation. Working with a single image makes disk mastering much more manageable. However, this convergence is also extremely welcome from a preservation point of view, as a single class of disk image can be used to cover a wide range of media.

For our older material, we must be able to cope with this variation in form, and even for newer materials, we need to be able to cope with the common variations in the way in which the media conform to the standards. This is particularly true for consumer writable media, where the software that creates the disks does not always behave reliably. This manifests itself not just as systematic deviation from the standards due to software or hardware problems, but also as variability in the quality of the disks due to the reliability of the creation process. For example, when creating ('burning') a writable CD, the process can fail and create unreadable disks (known as 'coasters'), particularly when the disk creation speed is high. For this reason, optical disks should be checked immediately after creation, but this is difficult to enforce when working with external parties.

With some assistance from the curators we were able to identify some particularly 'difficult' collections, and used those as a starting point to determine what type of variation there was in the optical media format. Across our collections, we encountered a very wide range of disk formats on optical media:

- DVD [8] or CD-ROM [9, 13] data disks in ISO 9660/UDF format (containing TIF, JPG, audio data files, etc.).
- DVD Video [8] disks in ISO 9660/UDF format [17, 18] (containing video data, e.g. VOB files).
- HFS+ (Mac) [3] format data disks.
- Red Book [14] Audio disks with sessions and tracks
- Yellow Book [13] Mixed-mode compact disks with a leading or trailing ISO 9660 data track containing mixed media alongside the audio tracks.
- Malformed 'audio' disks arranged in audio-like tracks, but the tracks themselves containing WAV files instead of raw CDR data.

The ISO 9660 specification [17] defines a disk image file format that can be used to clone data disks. This approach gives one single archive file that includes all the digital files contained on a CD-ROM, DVD, or other disk (in an uncompressed format) and all the file system metadata, including boot sector, structures, and attributes. This same image can be used to create an equivalent CD-ROM, and indeed mastering data disks is one of the purposes of the file format. It can also be opened using many-widely available software applications such as the 7-Zip file manager [15] or the WinRAR archive shareware [21].

Similarly, for later disks, such as DVDs, the UDF disk format specifies the layout of DVD disks and a general DVD image file format which is backwards-compatible with ISO 9660. The situation is similar for HFS+, as the data can be extracted as a single contiguous disk image without any significant data loss.

While CD-ROM, DVD and HFS+ format disks are reasonably well covered by this approach, there are some important limitations. For example, the optical media formats all support the notion of 'sessions' – consecutive additions of tracks to a disk. This means that a given carrier may contain a 'history' of different versions of the data. By choosing to extract a single disk image, we only expose the final version of the data track, and any earlier versions, sessions or tracks are ignored. For our purposes, these sessions are not significant, but this may not be true elsewhere.

For DVD disks, the main gap is that the format specification permits a copy protection system that depends on data that is difficult to capture in a disk image. Specifically, a data signal in the lead-in area of the disk contains information required to decrypt the content, but most PC DVD drives are unable to read this part of the disk, by design. Fortunately, this does not represent a problem for the Endangered Archives content, as it does not rely on media that use DRM or other technical right restriction methods.

The situation for Red and Yellow Book Compact Disks [13] is significantly more complex. As mentioned above, the overall disk structure, the sessions and tracks that wrap the data, are not covered by the ISO 9660 file format. Furthermore, it does not capture the additional 'subchannel' data that lies alongside the main data channel, which is used for error correction, copy protection and more esoteric purposes (see the CD+G standard [31] for an example). This information is often hidden from the end user, and indeed many CD drives are unable to access subchannel data at all.

Any attempts to preserve the full set of data channel, session and tracks is inhibited by the fact that there is no good, open and mature file format to describe the contents of a CD precisely. Proprietary and ad-hoc formats exist, but none are very widely supported, standardized or even documented. Even for simple Red Book Compact Disk Digital Audio media, there is only one standardized, preservation-friendly format that accurately captures the session, tracks and gaps – the ADL format [26]. This is a relatively new standard, and is not yet widely supported. Given this situation, we were forced to choose whichever file format is the most practical in terms of the data it retains, given the types of content we have, and by how well the tools that support those formats can be integrated with our workflow.

2.2 Disk Images Choices

Due to the variation in media formats outlined above, our overall migration workflow must be able to identify the different cases and execute whatever processing and post-processing steps are required. The first decision we must make, therefore, is to decide what type of disk image we should extract for each type of disk.

If the original data-carrier is a valuable artifact, data-carrier disk images should be produced and treated as the preservation copy.

Similarly, if file system metadata contained in a disk image may contain significant characteristics of the digital object that should be preserved (as is the case, for example, for bootable magazine cover disks) then the disk image should be treated as the primary preservation object. Ideally, this should be in a format that captures all of the data on the disks, not just the data from the final session.

In contrast, if the data-carrier could be considered simply a transfer medium and direct access to the data files is desired, they can be extracted as simple files instead. However, as indicated earlier, this can only be done once the data-carrier metadata has been properly evaluated, so practically we extract as full disk images at first, and then carefully generate the preservation master files from that image. Thus, we decided that all CD-ROM and DVD data or video disks should be ripped to ISO 9660/UDF disk images.

Similarly, the HFS+ disks should be ripped as single image bitstreams containing the volume data. These also manifest themselves as disks containing a single data track, but that happens to be HFS+ formatted instead of using ISO 9660/UDF. Therefore, the process of extracting the data track is identical to the previous case, and the difference lies only in the post-processing procedure.

Unfortunately, the Red Book, mixed-mode Yellow Book and malformed disks could not be extracted to ADL, as the available tools did not support that format. Those tools only supported the proprietary Media Description (MDS/MDF) file format (no public specification), which limits the range of post-processing tools we can use, but which can contain all the information on the disk and thus could be migrated to a format like ADL in the future.

For the Red Book disks, the content of the MDS/MDF disk image files can then be extracted in post-processing with extraction software such as IsoBuster [23]. Unlike the other extraction software we experimented with, IsoBuster could identify and read the full range of disk images we encountered, including HFS+ disk images. The breadth of formats supported was the main reason why IsoBuster was our preferred tool for post-processing MDS/MDF disk images.

Unfortunately, IsoBuster was not able to extract the audio track data reliably when operated in batch mode, and we found the most robust workflow was to use IsoBuster to migrate the disk image to another format with broader tool support. This second image format is known as CUE/BIN format (no public spec), and consists of a pair of files where the cuesheet is a simple text file describing the tracks and their arrangement on the disk, and the binary file contains the concatenated data from each tracks. This format is therefore less comprehensive that MDS/MDF, as the sessions and subchannel data have been discarded, but allows other software such as bchunk [11] to be used to produce usable WAV files from the raw binary data. The mixed-mode Yellow Book disks ripped in the same way as the audio disks, but extracting the content is slightly convoluted. After using bchunk to extract any ISO 9660 data tracks, each must be further processed to extract the files.

The malformed disks can also be ripped to MDS/MDF format, but complicate the content processing workflow further. After bchunk has been used to extract the tracks, they must be characterized using the 'file' identification tool [5] to see if they contain the RIFF header indicative of a WAV bitstream.

2.3 The Robot and the Automation Stack

The fundamental limitation on the throughput of the migration process is the manual handling process; the moving and cataloging of disks, and the opening and closing of jewel cases. Critically, the EAP disks are individually labeled by hand, were kept in sets associated with a particular project, and the ordering of the disks had to be retained. When processing the disks, the association between the physical item and the electronic image must be maintained, and so the overall workflow must ensure that the disk identifier is captured accurately and can be associated with the right disk image. The design of the media processing workflow took these factors into account and optimized to usage of the available staff effort while minimizing the risk of displacing or exchanging any disks.

Originally, we started working with a very large-scale disk robot: an NSM 7000 Jukebox (see e.g. [6]) fitted with 510 disk trays and 7 drives. While in principle such a large machine should allow high-throughput migration of optical media, there were a number of issues that made this approach unsuitable. Firstly, while the hardware was essentially sound, the accompanying software was intended for writing to a pool of disks, rather than reading a stream of disks. Trying to make the machine run 'in reverse' was extremely cumbersome, and such attempts were rapidly reduced to firing SCSI commands directly to the disk robot and ignoring the supplied software stack almost entirely. The details of the hardware design also worked against us. For example, the cartridges and disk trays used to load the machine had been optimized for storing sets of disks on shelves in the cartridge after the data had been written to them. This lead to a very compact physical design, but made the process of loading, unloading and reloading the cartridge with fresh disks rather awkward and error prone. In fact, all the robot solutions we looked at were primarily designed for the mass-write use case, but the NSM 7000 support for large-scale reading was particularly lacking.

Putting the media loading issue aside, we found that the main efficiency problem arose from the way exceptions, i.e. damaged, malformed or unusual disks were handled. If all the disks were perfect then the large-scale solution could be made to work fairly easily, with the operator loading up the machine and then leaving it to process a large number of disks autonomously and asynchronously (during which time the operator could perform other tasks). However, a small but significant percentage of the media we have seen have some sort of problem, and so the efficiency of the overall workflow is critically dependent upon this exception rate. This is because the task of reliably picking the problematic disks out of the whole batch rapidly becomes very difficult and error prone when the batches are large. It was, of course, of importance not to misplace or exchange any of the disks from the original collection, and a complex exception-handling process makes this difficult to ensure.

The British Library Sound & Vision group, in comparison, has successfully processed large amounts of compact disks using a single workstation with an array of 10 disk drives. This works well because, as all items are clearly distinct, individually catalogued and barcoded, they can be scanned and processed rapidly. As each disk represents an independent work, the fact that a manually loaded drive array will tend to process disks asynchronously (and thus not retain the disk order) is not a problem. Any problematic disks only occupy a single drive, and the others can continue to be loaded without blocking. Unfortunately, this approach was not well suited to our content, due to the order of batches of disks involved, and the manual cataloguing required per disk.

Following these experiences, we moved to using much smaller robots, the DupliQ DQ-5610 [1] and the Nimbie NB11 [2]. These small-scale machines are only capable of processing tens of disks, but by breaking the collections up into batches of manageable size, the exceptions can be handled more gracefully. This size limitation can then be overcome by running more than one machine in parallel, allowing the process to be scaled up quite effectively as each batch is processed independently. Any exceptions encountered can be tracked more easily, and brought together into a single, manually inspected batch.

Both units are USB 2.0 devices comprised of a single CD/DVD drive and a robotic component that handles the disks. However, the precise physical mechanism is different. For the DupliQ, a robotic arm grips and lifts the disks by the central hole, passing them from a lower tray to the drive and, once the disk has been processed, from the drive to an upper tray, giving a Last-In-First-Out (LIFO) processing order. The Nimbie has a simpler mechanism, with the disks held directly over the drive tray and released by a turning screw, leading to a First-In-First-Out (FIFO) processing order. In general, we found the Nimbie mechanism to be more reliable, as the DupliQ gripper mechanism would frequently fail to grip disks, could not cope with disks sticking together, would sometimes drop or even throw disks, and following this type of hardware error, the software would usually cope poorly and the batch would have to be restarted. Also, the DupliQ can only be loaded with about 25 disks, whereas the Nimbie can be loaded with up to 100 disks, and due to the FIFO ordering, can be run continuously if necessary.

Both small robots also came with appropriate software that supported extracting the disk data as disk images, called QQBoxxPro3. Unfortunately, this proprietary software also forced us to adopt the MDS/MDF format, as this is the only format it supports for multi-track/session disks. However, a more significant limitation was the lack of configurability for different types of disks, meaning that we could not instruct the robot to rip the contents of the different types of disks in different ways, as we would ideally like.

The DupliQ robot came with version 3.1.1.4 of the QQBoxxPro3 software, which appeared to assume that any single-track disks should be ripped as ISO 9660/UDF data disks (with an '.iso' file extension), whereas multi-track disks were ripped as MDS/MDF disk images. This was helpful for data disks and DVD Videos, but potentially quite dangerous for the HFS+ format disks, as that format is quite similar to the ISO 9660 format, and so some software tools open up the disk image and attempt to interpret it as ISO 9660 data without any warning. This makes is appear as if that data is corrupted and/or missing.

The Nimbie robot came with a later version of QQBoxxPro3 (3.3.0.5), which instead simply ripped all disks as MDS/MDF images. This leads to more complete disk images, but means that all of them require significant post-processing to access the data.

For our use pattern it turned out that using the DupliQ's version of the software with the Nimbie robot created the most effective configuration.

It is worth noting that the choice of disk-copying robot can depend on the composition of disk formats in the collection that were discussed in Section 2.1. We ran samples of difficult files on another FIFO robot, the MF Digital Data Grabber Ripstation [10], whose hardware is very similar to the Nimbie. The software on the 2 robots produced useful images for different disk carrier variants, ejecting disks in different situations and left differently useful log information that permitted identifying the presence of problematic situations. Depending on the expected distribution of disk file formats, one or the other of the two robots would have been preferable.

2.4 Disk Copying Workflow

Batches of disks were received from the curatorial group, and placed in a safe location next to the disk processing station. This station consisted of a basic PC with the USB robot, and a number of other items listed in Section 5 below. Large sets of disks were broken down into manageable batches of around 30 disks. Initially, this was because of the batch size restriction of the DupliQ machine, but due to the manual-handling overhead introduced by the problematic disks, this relatively small batch size was retained so that the exceptions could be managed effectively.

For sets of disks with low exception rates, the FIFO nature of the Nimbie machine proved very useful, as larger sets of disks could be continuously loaded into the machine. Of course, having multiple machines allows the processing of separate batches to be parallelized, and we found this to be a very effective approach. Initially, we set up two processing stations, running a DupliQ unit and a Nimbie in parallel. However, working with the two different disk-processing orders (LIFO v. FIFO) meant performing two different cataloging processes, one noting the disk identifiers in reverse, recording the disk metadata became a risky procedure. Furthermore, as indicated above, the DupliQ hardware was slightly less robust and reacted badly to difficult disks. Therefore, we moved to running two Nimbie units in parallel instead.

With these two processing stations running in parallel we were able to achieve processing rates of 1,050 disks per month, corresponding to data rates of 2.2 TB per month. The parallel robots had allowed us to minimize the fraction of the time spent waiting for the processing to finish, before the next load of disks could be processed, while overlapping the manual handling with the extraction process as much as possible. The limiting factor in the process at that point was the need to manually create metadata.

2.5 Handling Defective Disks and Other Exceptions

We found a range of particularly problematic disks, with the majority of them being physically malformed in some way that made them unreadable. In some cases, this manifested itself as disks that hung for long times in the drive, after which they were reported as being unrecognized. In others, the extraction would start as normal, but would slow down and eventually hang due to some local disk error. In rare cases, manual recovery of these disks was possible just using a different combination of drive hardware and ripping software. Usually, however, our only option was to make the curators aware of the issue as soon as possible, so that they could get in touch with the original authors promptly and get the content re-submitted on new media. In general, we found that disk problems or failures were correlated, i.e. most projects would have no problems, but some would have many problems. It was not possible to determine the root cause of these problems, but clearly systematic failures during the disk-burning process seems to be a more likely cause than simple disk aging or bad disk batches due to manufacturing defects.

Sometimes the cause for stabilization failures was unrelated to production properties of the disk or files. Problems with statically charged disks could be handled by attaching anti-static straps. Similarly, dirty disks tend to stick together. Since the EAP disks

are not valuable artifacts that need to be protected in the long run and since our disk drives were inexpensive and did not justify thorough disk cleaning in order to protect the drives, we did not rigorously clean disks as a matter of principle. If visual inspection showed dust we used a powerful camera lens cleaner to blow it off. A compressor proved to be too noisy for the shared office environment. More stubborn dirt was washed off using a solution of distilled water and isopropyl alcohol in equal parts. We used camera lens microfiber cloths for wet and subsequent dry cleaning and were careful to clean disks radially from the center of the disk straight to the outer edge in order to avoid inadvertently scratching consecutive data. We received valuable advice from the British Library Sound and Vision studios on disk cleaning issues. In one case, the cause of stabilization failure turned out to be labels that were affixed to the disks. The paper used for the labels was too thick and caught in the disk drive. We placed moist cloth onto the labels in order to soften and peel them off. For severely scratched disks we were able to borrow a disk polishing machine that physically polishes off some of the disk's thickness to remove scratches.

3 CONTENT MANAGMENT

Stabilization is only the first step in improving the preservation quality of the content submitted on hand-held data-carriers (see Table 1). Once the material is stabilized it experiences changes for curatorial, preservation or access reasons. This includes identification of duplicate, damaged or problematic images, securing replacement images, re-organization of content to standardize the collection structure, cataloguing, generating access surrogates, and the embodiment of these processes and relationships as archival metadata. It is striking that this process can extend over a long period of time (possibly years) due to the large volume of projects, projects continuing to submit materials, and the manual intensiveness of the curation.

Figure 1 illustrates the content management evolution that will be discussed in the following sections.

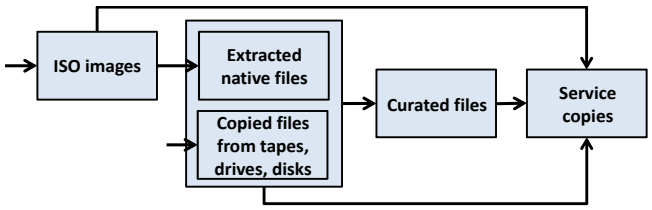


Figure 1: Overall content evolution, blending resources from different carriers into a curated, accessible artifact

This multi-stage workflow creates multiple versions of the same item, and the pressure this creates on the available storage space means that that previous versions need to be overwritten in order to keep costs down. This means that the process needs to be extremely well documented, so that ancestor items can be purged with confidence. Furthermore, the long timeframe of each project means that we must be able to add content over time, and our data management procedure must be able to cope with this. We must, therefore, diligently collect a range of metadata during data processing in order to ensure the content can be managed well over time.

For example, a fully stabilized data-carrier should consist of a set of files (disk image or content files) with associated metadata, linking the online content to the physical data-carrier via its identifier, linking to other metadata provided with the content, and documenting the stabilization event as part of the content's provenance. We chose to record all metadata in simple Excel spreadsheets that enforce controlled vocabularies and detect duplicate values when appropriate. When designing the spreadsheets we made sure that all fields could later on easily be translated into the final METS [28] and PREMIS [29] formats we tend to use. Further details of the metadata we chose to capture can be found in section 4.

3.1 Content Evolution

The overall content workflow shown in **Figure 1** is built up from a chain of more fine-grained processes and actions. In this section, we discuss some of those processes and the particular preservation issues that arose during their implementation.

3.1.1 Selecting the Data-Carriers to be Stabilized

Curators select the data-carriers that should be stabilized next based on prioritizing and balancing the following factors:

- risk assessment (e.g. age of the data-carriers, expected level of problems with reading the data-carrier, or expected difficulty of re-requesting material for faulty data-carriers),
- quick gain (process large external hard-drives over DVDs over even smaller CDs),
- user demand for the content.

If there are duplicate data-carriers (possibly at different resolutions) the curators should chose the highest quality copies and, if this is not obvious, ask digital preservation advice. Simple rules of thumb, such as preferring a TIFF file to a PDF, don't always apply. For example, we found low-resolution files with duplicate PDF's that contained higher-resolution TIFF duplicates.

3.1.2 Stabilize the Data-Carriers

The content migration itself, as described in Section 2, is embedded in a physical media management workflow. The disks chosen for stabilization are transferred to the stabilization station and the location of the data-carriers is recorded on a dedicated spreadsheet. This is usually a single large batch corresponding to a single project. This batch is then broken down into smaller batches, so that they can be processed efficiently. Physical place markers are used to keep track of each batches location in the relevant storage boxes. Processed disks are returned to this location, whereas failed disks are replaced with colored disk sleeves that contain a written record that identifies the disk, the nature of the failure and what has and is being done in order to resolve the failure. Stabilization metadata (as described in Section 4) is created as the robots are being loaded. Unidentified or inappropriately identified disks receive a unique data-carrier identifier which is recorded in the disk hub and the metadata spreadsheet. At this step, we also physically clean disks, should this be necessary.

3.1.3 Clean up Disk Image Directories

Clean-up of directories containing disk images can occur once stabilized batches of disk images are combined into one online project. It involves removing metadata, directory structures and files that were previously only needed to manage the stabilization process and merging and de-duplication of the batched files. We have decided to execute the latter process in a just-in-time fashion, i.e. when the following extraction step is to be executed, rather than every time a new batch is added.

At this point we can also execute the post-processing of disk images for the rarer data-carrier variants that was discussed in Section 2.2. If the disk-copying robots by default produced output formats that were not optimal for the data-carrier variant, this is

recognized at this point and the actually desired output disk image formats are created. For example, we create non-proprietary WAV/ADL files from proprietary MDS/MDF files. Again we need to create checksums for the newly created files.

3.1.4 Extract Content for Curation

In order to permit content curation to take place, an additional goal was to make data files individually and uniformly accessible. Disk images, tape and hard-disk representations received from the original suppliers are represented differently and require different modes of access. While the desire to preserve as authentically as possible drives us to store disk images, this form should be made transparent to the user so that they can quickly and reliably access the files themselves. Representing files universally in a uniform file-system and directory structure with uniform structural metadata makes it possible to access them all in the same way.

For disks we can use software tools such as IsoBuster [23] to extract the files from the disk images. The stabilization of digital objects from tapes or external hard drives best happens directly as files in their native file format since image formats and image extraction software for them are not in wide spread use (particularly for the MDS/MDF format, as noted earlier).

Where disk image formats are available, alternatively, data files can transparently be extracted on demand using software appropriate for the disk image.

3.1.5 Select and Create Curated Folders and Files

In the cases where data-carriers are valuable artifacts in their own right, the disk image is considered the preservation copy of archival value. No further curation of the content is desired.

In the cases, however, where the data-carrier mostly functioned as a means of data transfer, curators may change the content files. They may reorganize folder structures, move files into a more logical order helpful to human interpretation, rename files for better human consumption and link to those files from their catalogues. They may discard any unwanted directory structures, delete or replace unwanted files, duplicate files or files that don't satisfy collection guidelines. This is a manual process, and these curated files and folders have enormous value added compared to the files merely extracted from the data-carriers, and therefore must be considered the ultimate preservation copy.

In addition, individual files may be changed for preservation reasons. One may migrate files that are not in a preservation-worthy format (e.g. proprietary raw files that are not supported, password protected files, etc.) If this is done, then the original item is also kept unless we have a very high degree of confidence that the migration was as sufficiently complete.

3.1.6 Generate Service Copies

Once files in their native file formats are available one can derive service copies for user access. Curators select the files from which service copies should be created for presentation to the end-user and staff responsible for the service copies generate them from the preservation copies. While they should be backed up and managed with customary care, they are not part of the digital preservation cycle. They can be recreated from preservation copies if the need should arise.

3.2 Storage Issues

The large size and complexity of the content and its evolutionary nature led to several network storage issues.

Most importantly, the disk images had to be stored alongside the content extracted from hard-drive submissions, and the curated versions of the content for each project. Additionally, some projects submitted material on mixed data-carrier types. We managed the relationships between the collection parts that are stored as disk images and those that are stored in the native file system in two ways. We recorded their structural relationships in metadata and we used uniform directory naming conventions that indicate how the parts of the collection relate. For example, we may have a collection of files stored in the file system that is derived from a collection of disk images. This derivative relationship is recorded in the project status spreadsheet and also expressed through directory naming conventions. Its directory structure looks different from the situation where half of the collection is submitted on disks that were stabilized as disk images and half of the collection was submitted on external hard-drives that were extracted as files, therefore forming a sibling relationship. This required clear folder naming conventions to be decided and enforced.

In practice, cleaned up versions or curated folders can over-write earlier representations of the same content. For now, we have a cooling off period for each collection in order to determine whether any problems are likely to arise. After this digital preservation experts and curators together decide whether older representations should be deleted.

Content that evolves slowly over time is a challenge in the presence of a write-once digital repository. In this case, a reliable intermediate storage architecture needs to be determined.

3.3 File Management Issues

In addition to the content management steps described above one needs to perform a large number of common file management tasks. These include

- creating checksums (discussed in the following sub-section),
- copying large batches of files across the network,
- merging and de-duplicating batches of files,
- managing and keeping track of the various batch locations that were created due to limitations in the maximum size of an available storage unit (about 8TB),
- renaming and moving files and directories whilst recording the relationship to the old names and folder relationships,
- identifying file formats,
- validating files and discovering corrupt files.

The biggest problem we found is that there does not seem to exist a risk-proof tool-kit of stable, uniformly applicable tools that can be handled by non-technical staff. Ideally many of the file management activities should be in the hands of the curators rather than in those of IT personnel or digital preservation staff. But having to customize scripts, having to choose different software tools for quite similar situations, and not being able to mask “dangerous” flags in operating system commands currently still require specialist involvement.

As an example of this issue, the following sub-section discusses creation and management of checksums.

3.3.1 Create Checksum Manifests

All disk images (and indeed all content) should have a verifiable manifest of checksums created as soon as possible during the item’s lifecycle. This can then be used to ensure that no items have become corrupted or lost over time. Therefore, as soon as possible after the disk images were written to the external hard-drive, a full manifest of checksums was created for the complete hard-drive. The disk images were then transferred to backed-up

network storage, using the checksum manifest to verify the transfers. Note that we chose to use SHA-256 checksums [21], as these are the type used by the institution's archival store and so can accompany a digital item throughout its entire lifecycle.

One problem we encountered is that checksums produced by different content creators or by different software can contain different variants (such as checksums run in binary or text mode) but that this is not necessarily indicated correctly in the manifest. Similarly, different tools use different text encodings and/or folder separators in paths (i.e. forward slashes or Windows-style backslashes). Additionally, the checksums can be laid out in different ways (e.g. per file, folder or collection). We choose top-level collection manifests so that completeness can be managed along with integrity, but this mode of operation was not well supported by all tools.

Another, more serious problem was these tools were not sufficiently robust or user-friendly. A range of GUI programs do exist, but most were found to have some circumstances when they did not behave as expected, or were difficult to use. This is of critical importance when we wish the curators or content creators to build the manifest early in the lifecycle. The best tool we have found so far is the ACE Audit Manager Local Web Start Client [24]. This can be launched easily and has a reasonable user interface. Ideally, this tool might be extended to allow more formal data transfer and storage structures (e.g. bags as per the BagIt specification [19]) to be generated from sets of files with known hash sums.

4 METADATA

We managed three types of metadata pertaining to the data-carriers we stabilized. The primary and secondary metadata described below are of a descriptive nature and are of particular importance for supporting the curatorial work. The stabilization metadata is the metadata we produced in order to support the stabilization process.

4.1 Primary Metadata

Primary metadata is provided by the projects to describe the content and to describe the data-carriers and their structure.

4.1.1 Submitted Metadata Listings

Metadata listings are submitted by each project separately from the content data. Most of them are submitted as one or more separate Microsoft Word or Excel documents, sometimes in the form of an Access database or in the form of a paper document. The metadata listings are recorded upon receipt and linked to the data-carriers via the project and data-carrier identifiers.

There is also non-digital metadata such as handwritten notes inserted with or written on data-carriers.

4.1.2 Target Metadata

Curators catalogue all projects using the ISAD(G) and ISAAR(CPF) descriptive metadata standards down to archival file level, sometimes item level.

4.2 Secondary Metadata

Besides explicit listings and content descriptions that are identified by the projects as primary metadata, there are also additional PDF, Microsoft Word documents or JPG images mixed in with the digitized content that document the circumstances, location or people associated with the collection. There is also some descriptive metadata included in the set of content data, e.g. a handwritten note giving descriptive or manifest information, which has been scanned and, together with the content, included in the submission. We even found a file that contained the password for accessing the remaining files in the folder.

It is currently not possible to detect this born-digital secondary metadata automatically from within the content files; rather it can only be identified manually since it requires curatorial judgment.

4.3 Stabilization Metadata

Stabilization metadata are the metadata produced as a result of the stabilization process. Stabilization metadata are created in order

- to document the provenance of the content files to provide information on the content's authenticity and to enable problem solving if a tool involved in the stabilization process should have caused problems,
- to link the resulting content files to their original data-carriers: Disk images or extracted files should be linked to the original data-carrier so that it can be found if there are problems with the extracted content or if questions need to be directed to the project that has submitted the material,
- to link the content to its primary and secondary metadata.

Additionally it must, at later processing stages, record:

- if directory or file names are renamed in order to be able to link to previous versions,
- if the bit-representations of files or images are changed to document the degree of authenticity of the content,
- from which preservation copy, and how, a service copy is derived to document the degree of authenticity of the delivered content.

We documented the following types of stabilization metadata:

Data-carrier identifier. Data-carriers may be uniquely identified through such information as project number, accession numbers, box numbers, disk order within the storage box, unique data-carrier identifier within the project that were assigned by the project, cataloguing codes, etc. We set up the metadata spreadsheets to flag up duplicate data-carrier identifiers, recorded duplicate naming events and created disambiguating identifiers. Data-carrier identifiers enable us to link disk images or extracted files to the original data-carrier.

Metadata about where the data-carriers are located during the stabilization process and who is responsible for them. They may be in their regular shelf space, at the stabilization processing workstation, separated out for further investigation of failures, or with a curator.

Metadata about the stabilization event. This is provenance metadata that keep track of what processes the content has undergone. We record for each project which data-carrier has been stabilized and to what degree of success:

- When the data-carrier was stabilized,
- Who stabilized the data-carrier,
- What software and hardware was used to stabilize the data-carrier,
- Status of the stabilization event. This may be closed-Successful, closed-Manual clone, closed-Partial clone, closed-Failed, open-Partial clone, open-Failed, Not attempted,
- Primary metadata accompanying the data-carrier. This
    - may be transcribed into the metadata spreadsheet,
    - may be an image scan or photo of the accompanying documentation,
    - may be linked to a primary metadata register via the unique data-carrier identifier,

- File names of the output files of the stabilization process, i.e. the names of individual disk images,
- File extensions of the output files of the stabilization process,
- Administrative metadata to help manage the handling of the batch, such as the number of disks per stabilization batch, how many attempts have been made, run time of the stabilization, comments that explain special circumstances, etc..

Metadata recording the project status. This records which stages of the processing workflow the project has undergone (when and using which software and hardware) following stabilization. This includes the processes outlined in section 3.1.

Storage information. This records location and quantity of the stabilized content.

5 APPENDIX: STABILIZATION WORKSTATION REQUIREMENTS

Setting up a fully equipped stabilization workstation may include the following items. In each case we list the item, why it is needed; and the instances we used without any intention of endorsing the particular product.

5.1 Hardware Components

PC: For cloning disks with disk copying robots onto external hard-drives, editing metadata, check-summing; Dell and HP standard issue PCs with Microsoft Windows XP.

Disk-cloning robot: To automate disk-handling, Nimbie NB11 disk robot.

External hard-drives: To store cloned images before transfer to the server. As large as possible, formatted as required by the wider environment (i.e. NTFS for our Windows environment).

Server with large storage: For copying digital objects from external hard-drives, processing stabilized digital objects and for intermediate storage; Microsoft Windows Server 2003 R2, Standard Edition, Service Pack 2, Dual Core AMD Opteron.

Mass storage: For medium-term storage of digital objects before ingest into the permanent digital repository.

5.2 Software Components

Spreadsheet software: For capturing metadata; MS Excel.

Cloning software for robot: To clone disk images; QQBoxxPro3 v. 3.1.1.4.

Cloning software and software for extracting files from ISO images (by hand): For quality assurance of cloned image disks and for manual cloning of failed disks; IsoBusterPro 2.8.0.0, (7-zip, WinRar as back-up options).

Media player software: To check if the video disk images or the disks failed by the robot can be played; Real Media Player, VLC Media Player 1.1.4.

Check-summing software: For creating check sums in order to validate the integrity of cloned images or files when they are stored or transferred; FastSum 1.7.0.452[20], ACE Audit Manager (local WebStart client)[24].

File transfer software: For transferring and combining large batches of files; RichCopy 4.0, Windows Explorer, rsync.

Audit tool: For periodic checksum validation and for detecting duplicates; ACE Audit Manager (server version) [25].

5.3 Additional Workstation Components

Barcode scanner: To scan barcodes of already catalogued items in order to link to existing metadata. Not used for EAP.

High quality disk reader: To investigate failed disks; Plextor Blu-ray Disk Drive (PX-B120U) and standard PC disk drive.

Camera or scanner: For taking pictures of inserted or written information that accompanies the data-carriers. These images form part of the data-carrier metadata.

Tripod or camera stand: For holding the camera in place.

Lighting: For camera and for inspecting physical disk damage to disks.

Dust blower and/or oil free airbrushing compressor with nozzle: For cleaning disks; 1 Giottos GTAA1900 Rocket Air Blower, AB-AS18 Mini Piston type on-demand compressor for airbrushing, compressor hose, Sealey SA334 - Air Blow Gun, airbrush hose adaptor - 1/4"bsp female to 1/8"bsp male.

Workbench mat: For catching dust, to prevent it spreading in the office.

Microfiber cloths, wet and dry, liquid dispenser with 50% isopropyl alcohol, **50% distilled water:** For cleaning heavily soiled disks; Visible Dust Ultra Micro Fibre Cleaning Cloth.

Place Markers For marking batch start and end and problem areas within disk storage boxes; cut from plastic backing of folders, bookmarks.

Disk sleeves To mark places in disk storage boxes where failed disks have been removed; into them we insert labels describing reason for, date of removal, etc.; Compucessory CD Sleeve Envelopes Paper with Window.

CD/DVD marker, xylene free, with, ideally, water-based ink: To write data-carrier identifiers in disk hub; Staedtler Lumocolor CD/DVD Marker Pens Line 0.4mm Black 310 CDS-9 (which is alcohol based)

6 CONCLUSION

We have developed a robust workflow for stabilizing hand-held data-carriers onto online storage. For disks, we considered different approaches of parallelizing the copying process: a large-scale and 3 small-scale LIFO and FIFO disk-copying robots as well as asynchronous stacks of disk drives. We found that a small-scale FIFO disk-copying robot was the best tool for the given collection. We managed to set up a workflow that parallelized the copying to a point where the time needed for copying was balanced with the time needed for the manual tasks of disk handling and metadata creation, thus optimizing our use of the available staff member's time. We also adjusted the process batches so that problematic disks could be handled successfully without upsetting the smooth running of the workflow.

Whether disk images, content files in their native file formats or curated files are to be considered the ultimate preservation copy depends on the nature of the collection. In particular, it depends on whether the data-carrier itself or the curated folders represent the archival artifact. It also depends on the likelihood of having to go back to originals. It may be appropriate to, for example, keep disk images and curated files if storage considerations permit this.

However, the most important consideration when exploring the potential migration workflows was to let the choice of technical solution be driven by the development of the physical workflow and the nature of the content, rather than to choose the technical solution first and try to force the manual process to work around

it. We have adjusted the workflow to be as intuitive as possible and have sought to trim away any unnecessary steps. We documented the process as carefully as possible and validated the documentation by asking a new staff member to execute the workflow just from documentation. This has helped us to make the overall process robust and resilient enough that the curatorial team involved can take over the data-carrier migration process and proceed independently.

Most of the difficulties that arose during the development of this procedure originated from the suitability of the available software. For example, it is unfortunate that the disk-copying robots' software was not able to automatically recognize the data-carrier variants and perform different stabilization activities based on them. This meant that we had to write software to identify problematic situations where the wrong output format was produced and remedy the action in a post-processing step. Similarly, another significant obstacle to developing readily usable stabilization workflows was the lack of robust, intuitive file management software that can be handled by non-technical staff without the risk of inadvertently damaging the collections.

Data-carrier stabilization is a very time-consuming task that should be included in the planning for any project that includes hand-held data-carriers. A total volume of 100 TB requires 5 person years to stabilize at a rate of 20 TB/year. Obviously this number varies greatly with the type of data-carrier and the incidence of failed carriers. Staffing needs and the expenditure for stabilization workstations need to be included in business plans. Staff must be detail-oriented and systematic, but also flexible to respond to previously unencountered situations that require novel technical or pragmatic solutions. We had to process a surprisingly large number of projects before new types of workflow exceptions became rare.

Finally, while not a major concern during this project, other work on data-carrier stabilization must consider any copyright issues. We stabilized a collection for which we had the content owners' consent for copying. For other hand-held data-carrier collections at the British Library the copyright issues are difficult and the effort required to sort out permissions prohibits their stabilization.

7 REFERENCES

[1] Acronova Technology Inc. DupliQ. http://www.acronova.com/duplicator_dupliq_usb.htm.

[2] Acronova Technology Inc. DVD duplicator, DVD copier, dvd autoloader, CD ripper, CD copiers, DVD publisher, auto loading system, LightScribe Duplicator- Nimbie USB. http://www.acronova.com/blu-ray_cd_dvd_duplicator_publisher_nimbie_usb.htm.

[3] Apple Inc. Technical Note TN1150: HFS Plus Volume Format. 2004. http://developer.apple.com/library/mac/technotes/tn/tn1150.html.

[4] Arcadia. Arcadia Fund. http://www.arcadiafund.org.uk/.

[5] Christos Zoulas et al. The Fine Free File Command. http://www.darwinsys.com/file/.

[6] Data Archive Corporation. DISC DVD-7000 DVD Library. http://www.dataarchivecorp.com/disc-dvd-7000.htm.

[7] Digital Preservation Coalition. Digital Preservation for Forensics. http://www.dpconline.org/events/details/31-Forensics?xref=30.

[8] DVD Format/Logo Licensing Corporation. DVD FLLC - DVD Format Book. http://www.dvdfllc.co.jp/format/f_nosbsc.html.

[9] ECMA International. Standard ECMA-130 Data Interchange on Read-only 120 mm Optical Data Disks (CD-ROM). http://www.ecma-international.org/publications/standards/Ecma-130.htm.

[10] Formats Unlimited Inc. Ripstation DataGrabber. http://www.ripstation.com/datagrabber.html.

[11] Heikki Hannikainen. bchunk v1.2.0 - BinChunker for Unix / Linux. http://he.fi/bchunk/.

[12] IEC. IEC 60908 ed2.0 - Audio recording - Compact disc digital audio system. http://webstore.iec.ch/webstore/webstore.nsf/artnum/023623.

[13] IEC. ISO/IEC 10149:1995. http://standards.iso.org/ittf/PubliclyAvailableStandards/index.html.

[14] IEC. IEC 60908 ed2.0 - Audio recording - Compact disc digital audio system. http://webstore.iec.ch/webstore/webstore.nsf/artnum/023623.

[15] Igor Pavlov. 7-Zip. http://www.7-zip.org/.

[16] ISO. ISO/IEC 13346-2:1999 - Information technology -- Volume and file structure of write-once and rewritable media using non-sequential recording for information interchange -- Part 2: Volume and boot block recognition. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=29942.

[17] ISO. ISO 9660:1988 - Information processing -- Volume and file structure of CD-ROM for information interchange. http://www.iso.org/iso/catalogue_detail?csnumber=17505.

[18] ISO. ISO/IEC 13346-2:1999 - Information technology -- Volume and file structure of write-once and rewritable media using non-sequential recording for information interchange -- Part 2: Volume and boot block recognition. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=29942.

[19] J. Kunze. draft-kunze-bagit-06 - The BagIt File Packaging Format (V0.97). http://tools.ietf.org/html/draft-kunze-bagit.

[20] Kirill Zinov. MD5 checksum software for Windows. http://www.fastsum.com/.

[21] National Institute of Standards and Technology, FIPS PUB 180-3, FEDERAL INFORMATION PROCESSING STANDARDS PUBLICATION, Secure Hash Standard (SHS), Information Technology Laboratory, Gaithersburg, MD 20899-8900, October 2008, http://csrc.nist.gov/publications/fips/fips180-3/fips180-3_final.pdf

[22] RARLAB. WinRAR archiver, a powerful tool to process RAR and ZIP files. http://www.rarlab.com/.

[23] Smart Projects. ISOBuster | CD DVD Data Rescue software, featuring BD HD DVD. http://www.isobuster.com/.

[24] The ADAPT Project. Ace:Webstart Client - Adapt. https://wiki.umiacs.umd.edu/adapt/index.php/Ace:Webstart_Client.

[25] The ADAPT Project. Ace:Main - Adapt. https://wiki.umiacs.umd.edu/adapt/index.php/Ace.

[26] The Audio Engineering Society. AES31-3 AES standard for network and file transfer of audio — Audio-file transfer and exchange. Part 3: Simple project interchange. http://www.edlmax.com/AES31.htm.

[27] The British Library. The Endangered Archives Programme. http://eap.bl.uk/.

[28] The Library of Congress. Metadata Encoding and Transmission Standard (METS). http://www.loc.gov/standards/mets/.

[29] The Library of Congress. PREMIS - Preservation Metadata: Implementation Strategies. http://www.loc.gov/standards/premis/.

[30] Wikipedia. Rainbow Books - Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Rainbow_Books.

[31] Wikipedia. CD+G - Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/CD%2BG.

Towards an Integrated Media Transfer Environment:

A Comparative Summary of Available Transfer Tools and Recommendations for the Development of a Toolset for the Preservation of Complex Digital Objects

Antonio Ciuffreda[1]
David Anderson[2]
Janet Delve
Leo Konstantelos
Dan Pinchbeck
School of Creative Technologies
University of Portsmouth
+44 (0)2393845525
[1] antonio.ciuffreda@port.ac.uk
[2] david.anderson@port.ac.uk

Winfried Bergmeyer[3]
Andreas Lange[4]
Computerspiele Museum
+49 3031164470
[3] bergmeyer@computerspielemuseum.de
[4] lange@computerspielemuseum.de

Vincent Joguin[5]
Joguin S.A.S.
+33 (0)457931226
[5] vincent@joguin.com

ABSTRACT

Efficient media transfer is a difficult challenge facing digital preservationists, without a centralized service for strategy and tools advice. Issues include creating a transfer and ingest system adaptable enough to deal with different hardware and software requirements, accessing external registries to help generate accurate and appropriate metadata, and dealing with DRM. Each of these is made more difficult when dealing with complex digital objects such as computer games or digital art. This paper presents the findings of several studies performed within the KEEP project, where numerous open-source and commercial media transfer tools have been evaluated for their effectiveness in generating image files to be integrated into an emulation-based preservation solution for complex digital objects. We provide suggestions for assembling toolsets based on the specifications of these tools, therefore providing a valuable source of information for media transfer activities.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Performance evaluation (efficiency and effectiveness).

H.3.7 [**Digital Libraries**]: Systems issues.

General Terms

Documentation, Performance.

Keywords

Digital Preservation, Image File, Transfer Tools, Optical Media, Magnetic Media, Digital Objects, Emulation.

iPRES2011, Nov. 1–4, 2011, Singapore.

1. INTRODUCTION

The continuous development of digital storage media in recent decades has caused serious problems for accessing digital data stored on deprecated media carriers [1]. As media carriers cease to be supported by computer manufacturers and therefore become obsolete, the data stored on them need to be transferred to supported storage media in order to remain accessible. This problem is greatly amplified in libraries and other memory institutions, where a large number of materials are stored digitally [2].

Media transfer gives rise to numerous challenges. Issues include creating a transfer and ingest system adaptable enough to deal with different hardware and software requirements [3], accessing external registries to assist with the generation of accurate and appropriate metadata, and dealing with Digital Rights Management [4].

In this context, the main findings of different studies carried out as part of the KEEP project[1] (European Commission, ICT-231954) are gathered and presented. One of the main outcomes of KEEP is a media Transfer Tool Framework (TTF)[2] designed to support intelligent and robust transfer of complex digital objects within a preservation workflow. Although KEEP is focused on emulation-based preservation strategies the TTF is equally applicable to a migration-based approach.

A pre-requisite for building the TTF was to assess available open-source and commercial transfer tools with a view to incorporating them in the framework. Transfer tools for both magnetic (3.5'' and 5.25'' floppy disks) and optical (CDs and DVDs) media

[1] The KEEP project is co-financed by the European Union's Seventh Framework Programme for research and development. For further information on KEEP see: http://www.keep-project.eu

[2] http://www.keep-project.eu/ezpub2/index.php?/eng/Work-packages/WP1-Transfer-tool

carriers were examined. The choice of these media carriers was determined by the scope of the KEEP project. Effectiveness in generating computer game images ready for integration into an emulation-based preservation solution was the main criterion for our analysis.

In the following sections a concise analysis of tools for magnetic media and optical media is provided. For each of these tools an outline of their software and hardware requirements is given, together with their reading configuration settings and image generation capabilities. A general overview together with a presentation of the test outcomes completes the analysis. A conclusive summary of the analysis performed is then offered, together with a set of recommendations for anyone intending to assemble toolsets based on the specifications and testing of these analyzed tools.

2. EVALUATION OF TRANSFER TOOLS FOR MAGNETIC MEDIA CARRIERS

This section presents the analysis of three transfer tools for magnetic media carriers. The list includes two commercial transfer tools, Disk2FDI and Catweasel, and a freeware open source transfer tool application named Nibtools. The tests on Disk2FDI were conducted with the assistance of its creator (and KEEP technical lead) Vincent Joguin whose specialist knowledge and experience helped produce optimal tool performance. No qualified expert assistance was provided during the test on Catweasel and Nibtools. This together with the use of sector-dumped formats which are less able to capture protections or custom formats may have been contributory factors in the higher failure rate noted for these two magnetic media transfer tools.

A total of 2268 floppy discs of 3.5'' or 5.25'' size were tested in the evaluation study of Disk2FDI. As Table 1 shows, the environments of these floppy discs varied and included the following: Amiga, Apple II, Apple IIGS, Apple Macintosh, Archimedes, Atari ST, BBC Micro and C64. The following emulators were used to determine the success of the reading process of the Amiga, Archimedes, Atari ST, BBC Micro, C64 and Apple-based floppy disks respectively: WinUAE[3], Arculator[4], Hatari[5], B-EM[6], Hoxs64[7] and CiderPress[8].

The evaluation study on the Catweasel and Nibtools transfer tools (see Table 1) was based on the reading process of the same set of floppy disks. Nine and seven floppy disks containing video games for the Amiga and C64 computers respectively were used for the Catweasel transfer tools. Due to inability to read image files from Amiga, only the seven floppy disks containing C64 video games were tested on Nibtools. In order to determine the success of the reading process of Amiga floppy disks the WinUAE emulator was used to render the produced image files. The Hoxs64 and CCS64[9] emulators were used instead for testing the image files generated by C64 floppy disks.

[3] http://www.winuae.net

[4] http://b-em.bbcmicro.com/arculator

[5] http://hatari.berlios.de

[6] http://b-em.bbcmicro.com

[7] http://www.hoxs64.net

[8] http://ciderpress.sourceforge.net

[9] http://www.ccs64.com

2.1 Disk2FDI

Disk2FDI[10] is a commercial tool produced by Vincent Joguin primarily for generating Formatted Disk Image (FDI) files from floppy disks. In addition to the Formatted Disk Image format, Disk2FDI provides support for the following formats: Amiga Disk File, Commodore 1541, Macintosh DiskCopy 4.2, Apple II DOS-Ordered, IBM FM (single density), IBM MFM (PC and many others), and Atari ST Disk Image. The main features of Disk2FDI require a custom, but very simple, 2-wire Disk2FDI cable.

The FDI format is able to represent precisely information stored on a floppy disk, including possible copy protections and non-standard disk layouts. This accuracy of data representation is achieved by capturing at a very low level the magnetic flux transitions on the floppy disk surface. Disk2FDI captures the signals emanated by the electric pulses as soon as these are generated by the floppy disk drive from the magnetic surface of the disk, and before they are further processed by the floppy disk controller.

FDI files are currently supported by the following programs: WinUAE and E-UAE (Amiga emulators), CiderPress (Apple 2 disk image management tool), B-Em (BBC Micro emulator), Arculator (Acorn Archimedes emulator) and Hoxs64 (Commodore 64 emulator). The complete specification of the FDI image file format is open and available as a text file (FDISPEC.TXT) within the Disk2FDI distribution archive (including the freely-downloadable trial version).

2.1.1 Requirements and Specifications

Disk2FDI requires a pure DOS operating system (including FreeDOS), and a DOS-accessible mass-storage device large enough to contain image files. Further requirements include: any processor from the Pentium family or any other faster processor, 32 MB (min.) RAM, a parallel port configured for DOS (ideally from a PCI parallel port card), a Disk2FDI cable, and a floppy disk drive attached to the motherboard (not through a USB interface).

2.1.2 Test Results

Out of the 2268 transferred discs tested only a single protected floppy disc from the Amiga family was found to be imaged incorrectly by Disk2FDI. The remaining floppy discs were read successfully and as result fully working FDI image files were produced.

The tests showed that the average reading time for an FDI image is approximately eleven seconds for each track, typically resulting in transfer times in the range of one hour per disk. Although FDI files are compressed using lossless algorithms, they are generally very large. For example a 3.5'' 1.44 MB floppy disk produces a 14 MB FDI file.

2.2 Catweasel

Catweasel[11] is a commercial floppy disk controller that provides access to a large variety of floppy disk formats thus enabling its users to perform a range of different tasks such as reading, writing and erasing contents of floppy disks. Catweasel uses a PCI card which communicates directly with the floppy disk drive.

[10] http://www.oldskool.org/disk2fdi

[11] http://www.jschoenfeld.com/products/catweasel_e.htm

Table 1. Results obtained from the tests performed on transfer tools for magnetic media.

System	Discs read	Environment	Format	Working images	Defective images	Success Rate
Disk2FDI	2268	Amiga	FDI	1486	1	99.95%
		Apple II	FDI	51	0	
		Apple IIGS	FDI	9	0	
		Apple Macintosh	FDI	11	0	
		Archimedes	FDI	48	0	
		Atari ST	FDI	492	0	
		BBC Micro	FDI	24	0	
		C64	FDI	146	0	
Catweasel	16	C64	D64	4	3	31.25%
		Amiga	ADF	1	8	
Nibtools	7	C64	G64	4	3	57.14%

In addition to the PCI card, the Catweasel manufacturer, Individual Computers, provides drivers for the Windows 2000, Windows XP, Linux and Amiga OS4 operating systems and an application named *Imagetool*[12] in order to perform a variety of tasks (i.e., writing, reading and erasing) on the floppy disk. Jumpers together with the necessary cables to attach the card to one of the floppy disk drivers and to the on-board floppy disk controller are also provided.

The process of capturing the magnetic flux transitions on the floppy disk surface takes place at a low level: the tool acts as a flexible floppy disk controller by directly reading (but also writing in this case) signals from (and to) the floppy disk drive. Catweasel also makes use of sophisticated algorithms to provide automatic error corrections of stream of pulses which are found off-center in corrupted or sensitive floppy disks.

Catweasel supports by default a wide range of disk formats from the 3.5'' and 5.25'' families such as Atari, Apple II, Apple Macintosh, MS-DOS and different Commodore disk drive series (Leighton 2008). Further disk formats can be handled by downloading drivers from the Internet and reprogramming the PCI card.

2.2.1 Requirements and Specifications

Catweasel requires a Windows (98SE, ME, 2000 or XP) or any Linux or Amiga operating systems. Any floppy disk drive (usually the 3.5'' and 5.25'' disk drives) required to support the desired disk formats previously mentioned can be used. The only floppy drives which are currently known to be incompatible with Catweasel are from the Mitsumi D359 series, Teacdrives with integrated flashcard reader, Citizen drives for Compaq computers and the Samsung SFD-321B. A maximum of two floppy disk drives can be attached simultaneously to this disk controller.

If the computer already has a floppy disk drive and an on-board floppy disk controller, the PCI card (to be inserted in a PCI bus slot) provided by Individual Computers can be attached directly to the floppy disk drive on one side and to the on-board floppy disk controller on the other side using the cables provided. In the absence of an on-board floppy disk controller, the communication between this and the PCI card can be ignored. If the drives are for 5.25'' or 8'' floppy disks, special adapters will be needed for attaching these drives to the PCI card.

Once the Catweasel card is attached and the driver installed, reading/writing operations can be performed via Imagetool. Catweasel offers three different types of sector-dumped images to choose from: *plain image file*, *d64 (with error info)* and *atr*. However, Catweasel does not provide a disk imaging facility to an accurate and generic format (such as the FDI format).

If the *plain image file* option is selected and the *Read Disk* button is clicked, Catweasel will generate a plain image file of the specified disk. The extension of the image file depends upon the specific floppy disk format. If the *d64 (with error info)* option is selected, Catweasel will generate an extended D64 image file providing sectors error information in addition to disk content. The selection of the *atr* option instead will produce an Atari disk image file.

2.2.2 Test Results

A test set of sixteen floppy disks containing Commodore64 (5.25'') or Amiga (3.5'') video games were used with Catweasel during the testing process. As it can be seen from Table 1, Catweasel read only five disks correctly, thus producing image files which could be accessed via the previosuly mentioned emulators. The remaining eleven floppy disks instead led to the generation of image files which could not be rendered correctly.

During our experiments the average time needed to generate an Amiga disk file with Catweasel was 60 seconds. Based on its official Webpage of this tool 50 seconds is required to read a 3.5'' Amiga 1760 kB floppy disk and to generate an image file from it.

2.3 Nibtools

Nibtools[13] is a free open source program for PCs for generating image files, modifying the formats of these image files and writing content to floppy disks from a defined set of CBM64 floppy disk drives. We consider the source data processed by Nibtools less accurate than the source data produced by Disk2FDi and Catweasel, as the capture of the magnetic flux on the floppy

[12] http://siliconsonic.de/t/catweasel-usermanual.pdf

[13] http://c64preservation.com/nibtools

disk surface in this tool takes place at a higher level: while Catweasel captures and processes directly the signal generated by the electric pulses from the disk drive, Nibtools captures and processes the generated stream of bits only after the original signal has been processed digitally to the GCR (Group Code Recording) format by the floppy disk controller. However, Nibtools is able to create more accurate images (to the CBM-specific G64 format) than the Catweasel software. Nibtools supports only two CBM64-related image files: G64 and D64.

2.3.1 Requirements and Specifications

Nibtools requires a Windows (NT, 2000, XP, Vista or 7), any Linux or a MS/DR/Caldera DOS (with CWSDPMI) operating system. Under Windows or Linux, OpenCBM 0.4.2 (or any other higher version) should be used. Nibtools requires a Commodore 64 floppy disk drive from the 1541, 1541 II or 1571 series. A XP1541 or a XP1571 parallel cable with a serial cable from the x-series (X1541, XE1541, XA1541 or XM1541) or a XEP1541, XAP1541, or XMP1541 combination cable are also required.

The CBM64 floppy disk drive needs to be attached to a parallel port of the PC using a suitable parallel and serial cable. A combination of cables can be used otherwise, although the use of parallel port add-ons for the floppy disk drive is preferred in order to achieve higher communication speed. In addition to these physical cables, the OpenCBM kernel device driver will be needed in Windows or Linux in order to permit connection between the floppy disk and the computer[14].

Reading and writing tasks in Nibtools are performed in a command line interface. Once the floppy disk drive has been connected, the command *nibread [options] filename.nib* will generate a NIB image file. An extensive range of options is available, including choices for specifying the floppy disk drive unit or the starting and ending track of the floppy disk or for performing (crude) reading verifications.

Once the NIB image file has been generated the command *nibconv filename.nib filename.yy* can be used to convert the NIB file into an image file of a format specified by the file extension in *filename.yy*. The file can be converted to a G64 image file or a D64 image file.

2.3.2 Test Results

Out of the seven floppy disks containing Commodore64 video games only four floppy disks were read correctly producing image files which could be accessed correctly via fully working emulators. The remaining three floppy disks generated image files which could not be rendered correctly.

The tests performed as part of the KEEP Project suggests an average of sixty seconds to generate a C64 disk image files.

3. EVALUATION OF TRANSFER TOOLS FOR OPTICAL MEDIA CARRIERS

The section presents the analysis of five transfer tools for optical media carriers. We tested four commercial transfer tools: Alcohol 120%, Daemon Tools, CloneCD and Blindwrite. Additionally we examined a free program transfer tool application named ImgBurn. Thirteen CDs or DVDs containing video games for the Windows 95/98, Windows 3.1 or the DOS environment were used (see Table 2) for each of the aforementioned transfer tools. The four discs used for Windows 95/98 had copy protection schemes of different types (TOC, ProtectCD VOB and SecuRom). In order to determine the success of the reading process of these discs the DOSBox[15] emulator was used to render the image files of the video games for the DOS and Windows 3.1 environment. A computer with Windows 95 environment was instead used for testing the produced image files of video games for Windows 95/98 environment.

3.1 Alcohol 120%

Alcohol 120%[16] is a commercial application for optical disc authoring and disk image emulation providing image file generation capabilities from CDs and DVDs. Alcohol 120% provides support for a wide range of CD (CD-DA, CD+G, CD-ROM, CD-XA, VideoCD, Photo CD) and DVD (DVD-ROM, DVD-Video, DVD-Audio) formats. One of the most interesting features of Alcohol 120% is the ability to bypass several copy protection schemes, such as SafeDisk, SecuROM and Data Position Measurement (DPM) and to create image files of PlayStation and PlayStation 2 file systems. Due to legal issues Alcohol 120% does not generate image files from DVDs with CSS protection.

3.1.1 Requirements and Specifications

In order to run Alcohol 120% any Intel/AMD-based PC with a Windows (2000, XP, Server 2003 or Vista) operating system is required. At least 32MB of RAM and 10GB of free hard disk space, one or more CD-ROM or DVD-ROM drives and one or more CD/DVD recorders are also needed. If more than 2 CD recorders are installed, 700MHz CPU and 128MB RAM are recommended. A CD/DVD recorder can be used as a reader if there is sufficient hard disk space to store a whole CD/DVD image.

In order to communicate with the required hardware devices Alcohol 120% uses the SPTD (SCSI Pass-Through Direct) device driver. Alcohol will be able to use any disk drive no matter what type of hardware interface is used to connect the peripheral device with the computer.

Depending on the type of disk inserted, Alcohol 120% provides a wide list of different data types including NormalCD, NormalDVD, PlayStation 2, SafeDisk and SecuROM. Alcohol 120% provides context-sensitive image file formats to choose from, such as Media Descriptor, CloneCD, CDRWin and Standard ISO, depending on the image data type selected. Users can configure the reading process by accessing options such as Skip Reading Errors, Fast Skip Errors Block, Advanced Sector Scanning, Reading Sub-channel Data from Current Disk and DPM; this last option is used for reading CD/DVDs with DPM protection mechanism. Alcohol 120% can read this mechanism and encode it into a Recordable Media Physical Signature (RMPS), which reproduces the effects of a DPM. The RMPS and other metadata of the original disk are then stored in a Media Descriptor (MDS) file.

If the image format option for the image file is set to a *Media Descriptor Image*, Alcohol 120% generates an MDF file (the Alcohol 120%'s proprietary disk image format) which includes the actual data on the disk, and a Media Descriptor Image file, containing the information related to the header and track of the disk.

[14] http://c64preservation.com/files/nibtools/readme.txt

[15] http://www.dosbox.com

[16] http://www.alcohol-soft.com

Table 2. Results obtained from the tests performed on transfer tools for optical media.

System	Discs read	Environment	Format	Working images	Defective images	Success Rate
Alcohol120%	13	Windows 95/98	MDS	3	1	92.3%
		Windows 3.1	ISO	3	0	
		DOS	MDS/ISO	6	0	
DaemonTools	13	Windows 95/98	MDS	1	3	76.92%
		Windows 3.1	ISO	3	0	
		DOS	MDS/ISO	6	0	
CloneCD	13	Windows 95/98	CCD	2	2	84.61%
		Windows 3.1	ISO	3	0	
		DOS	ISO	6	0	
Blindwrite	13	Windows 95/98	B6T	3	1	92.3%
		Windows 3.1	ISO	3	0	
		DOS	B6T/ISO	6	0	
ImgBurn	13	Windows 95/98	ISO	0	4	69.23%
		Windows 3.1	ISO	3	0	
		DOS	ISO	6	0	

If the image format option for the image file is set to a *CloneCD Image File* Alcohol 120% will create an IMG file containing the data on the disk, a SubChannel Data file which stores the sub-channel data from all the tracks of the disk and finally a CloneCD Image file, which contains information associated with the logical structure of the disk.

If the image format option for the image file instead is set to a *CDRWin File*, Alcohol 120% will generate a BIN file containing the data on the disk and the CDRWin file containing track information. The selection of the *ISO Image File* option will cause Alcohol 120% to create a single ISO file containing the entire data content of the disk.

3.1.2 Test Results

All but one of the 13 test discs was read correctly and fully working MDS or ISO image files were created as result. The CD protected with the VOB ProtectCD produced an MDS image file which could not be accessed via emulation.

Reading speed configuration options are available only for CDs but not for DVDs. The maximum reading speed available during the image creation process for a CD was 24x (around 3.600 MB/sec), although the highest reading speed recorded with this optical medium was 12x (1.848 MB/sec). The highest reading speed recorded with a DVD instead was 11.2x (15.512 MB/sec).

3.2 Daemon Tools

Daemon Tools[17] is a commercial optical disc authoring and disk image application offering image file creation from disks inserted in CD, DVD, HD-DVD and Blu-ray drives, in addition to other related functionalities such as burning, converting and editing image files and erasing disk content.

[17] http://www.daemon-tools.cc/eng/products/dtproAdv

3.2.1 Requirements and Specifications

In order to run Daemon Tools a Windows (2000, XP, 2003, Vista or 7) operating system is required. A minimum of 500 MHz CPU, 256 MB of RAM is also required, in addition to one or more CD-ROM or DVD-ROM drive.

Daemon Tools uses the SPTD (SCSI Pass-Through Direct) device driver to access the required hardware devices. Daemon Tools can be used with any disk drive, regardless of the hardware interface used to connect the drive with the computer.

Irrespective of the nature of the CD or DVD, the following output file formats can be chosen: *MDS/MDF*, *MDX* and *ISO*. Depending on the image file format chosen and on the data type selected, Daemon Tools provides a list of additional settings for configuring the reading process. This includes settings for choosing a desired Data Position Measurement reading speed, for choosing the number of times the application can retry reading a specific disk sector if it is damaged and ignore any faulty sector found during the reading process. Daemon Tools also provides a list of configuration settings related to the generated image. This includes settings for reducing the size of the generated image file and for securing the generated file via a password-based security mechanism.

If the image format option is set to *MDS/MDF* two distinct files will be created: an MDF file, which includes the actual data of the disk, and a Media Descriptor Image file which contains the set of information related to the header and tracks of the disk. If the image format option instead is set to *MDX* an Extended Media Descriptor file, together with a MDF file and the Media Descriptor Image file previously mentioned will be generated. If the image format option instead is set to *Standard ISO* a single ISO file containing the entire content of the disk will be generated.

3.2.2 Test Results

In the test three copy-protected CDs resulted in defective image files which could not be accessed via emulation. The remaining ten discs were read correctly and fully working ISO or MDS image files were generated.

Reading speeds for both CDs and DVDs are customizable. The maximum reading speed available for a CD is 24.0x (3600 KB/sec), while the maximum reading speed for a DVD is 8.0x (10820 KB/sec).

3.3 CloneCD

CloneCD[18] is a commercial optical disc authoring application which can create image files from CDs or DVDs. CloneCD can read CD-R, CD-RW, DVD, Multisession and Digital Audio (CD-DA) disks. CloneCD provides support for DVD split file image formats, in addition to ISO and UDF formats. It also bypasses the SafeDisk 3 protection scheme during the reading stage.

3.3.1 Requirements and Specifications

In order to run CloneCD a Windows (98, ME, 2000, XP, Vista or 7 of 32 or 64 bit) operating system is required. An IBM-compatible personal computer with a 500 MHz Pentium microprocessor or higher, at least 64 MB of RAM is also needed.

CloneCD makes use of Elby CDIO, a novel device driver developed by Elby, the manufacturer company of this disk authoring tool. The widespread ASPI (Advanced SCSI Programming Interface) interface is not supported by this program though. Any disk drive, regardless of the hardware interface used to communicate with the computer, can be used with this authoring application.

CloneCD does not offer any choice of data formats for the image file to be generated. For CDs, CloneCD provides the following profiles for selection: Audio CD, Data CD, Game CD, Multimedia Audio CD and Protected CD Game. If the disk inserted is a DVD, the application offers only the DVD profile.

If the disk inserted is a CD the application will generate three files: an IMG file which contains the actual data of the disk, a Channel Data file containing the sub-channel data from all the tracks of the disk and a CloneCD Image file containing information of the logical structure of the disk. If the disk inserted is a DVD two distinct files will be created instead in the folder specified: a DVD file which includes the raw data of the DVD disk and a single ISO file containing the data content of the disk.

3.3.2 Test Results

Of the thirteen discs used during the testing process eleven discs were read correctly and fully working image files were generated by these discs. The remaining two copy-protected discs (ProtectCD VOB and SecuRom) led to the generation of image files which could not be rendered via emulation services.

CloneCD does not provide a configurable reading speed. The maximum reading speed recorded for a CD was 11.99x (2110 KB/sec). The maximum reading speed recorded for a DVD was 5.02x (6777 KB/sec).

3.4 Blindwrite

Blindwrite[19] is a commercial authoring program that allows image files to be created from CD, DVD and Blu-Ray disks. Blindwrite handles every available CD format, most of the DVD formats (DVD-R, DVD+R, DVD-RW, DVD+RW, DVD-RAM, DVD+R Double Layer and DVD-R Dual Layer) and Blu-ray formats (BD-R and BD-RE). Disks for game consoles such as Xbox, Wii and Playstation are also supported.

3.4.1 Requirements and Specifications

In order to run this authoring tool the Windows (XP, Vista or 7) operating system is required. Blindwrite needs an Intel Pentium III or an AMD Athlon or any higher processor. At least 512 MB of RAM with Windows XP or 1 GB of RAM with Windows Vista and 4.3 GB of available hard disk space are also required.

Blindwrite uses the Patin-Couffin CD device driver in order to provide access to CD, DVD and Blu-ray drivers. Any disk drive can be used with this tool, regardless of the hardware interface used to connect the peripheral device with the computer.

Blindwrite offers eight profiles to choose from: *Automatic*, *Audio CD*, *Audio CD+G*, *Bad Sectors* (for detecting automatically bad sectors in the disk), *ISO Image*, *No Split*, *Normal* and *Nibble*.

Regardless of the type of disk inserted, Blindwrite will generate the two following types of files, unless the ISO profile has been selected: a Blindwrite 6 Track Information file, a file type associated primarily with this application, which contains information of the tracks, and a Blindwrite 6 Disk Image file, which contains the actual data of the disk. If the ISO profile has been selected, an ISO file will be created along with the previously mentioned Blindwrite 6 Track Information file.

3.4.2 Test Results

During the testing phase a single CD with the ProtectCD VOB V.5 protection scheme led to the creation of an image file which could not be emulated. The remaining twelve discs were read correctly and fully working B6T, B5T or ISO image files from these discs were produced.

Blindwrite provides configurable reading speed. The maximum reading speed provided is 48x (7 MB/sec) for CDs, 16x (21.6MB/sec) for DVDs and 8x (36 MB/sec) for Blu-ray disks. The reading speed during the image generation however could not be evaluated as no information was provided by the application.

3.5 ImgBurn

ImgBurn[20] is a free authoring program for CDs, DVDs and Blu-ray disks providing image generation from disks and from files on the host computer or on a network, in addition to image writing and full disk readability verification. The main drawback with this program is the inability to read sub-channel data from a CD. ImgBurn does not provide support for multi-session disks and raw disks for disk burning activities.

3.5.1 Requirements and Specifications

ImgBurn requires a Windows (95, 98, Me, NT4, 2000, XP, 2003, Vista, 7 or 2008 R2) operating system. ImgBurn also supports Wine[21], thus allowing users to run this application on Unix-like operating systems. A 1.7 GHz Pentium IV processor or higher and at least 512 MB of RAM is also needed to run this authoring tool.

[18] http://www.slysoft.com/en/clonecd.html

[19] http://www.vso-software.fr/products/Blindwrite/blindwrite.php

[20] http://www.imgburn.com

[21] http://www.winehq.org

ImgBurn supports the following device drivers in order to access the external disk drives: ASPI (WNASPI32.DLL), ASAPI (ASAPI.DLL), SCSI Pass Through Interface, ElbyCDIO and Patin-Couffin. ImgBurn will work with any type of disk drive, regardless of the hardware interface used.

Independent of the type of disk ImgBurn provides the user with the following list of file formats to choose from: Bin Files, IMG Files and ISO Files. A list of configuration settings related to the image reading process is also provided. Settings include editing the maximum size limit of the generated image file, configuring the reading modality of the inserted disk, setting the number of retries after the reading failure of a disk sector and skipping the disk sectors that can't be read.

If the image to be generated is a BIN file and the inserted disk is a CD ImgBurn will generate an IMG file, a CDRWin file containing the information of the disk header and tracks and a CloneCD Image file, if this has been selected in the Settings section of the program. If, in this context, the user has decided to not generate any layout image files, ImgBurn will create a BIN file and a CDRWin file. If the inserted disk is a DVD, the program will create a BIN file including the actual data on the disk in addition to DVD and/or Media Descriptor Image layout files, provided that these have been selected in the Settings section. If the image to be generated is an IMG file and the inserted disk is a CD, two files will be created: an IMG file containing the entire content of the disk and a CDRWin file containing the information of the disk header and tracks. If the inserted disk is a DVD, an Image file will be generated, in addition to a DVD and/or Media Descriptor Image layout files, provided these two files have been selected in the Settings section. If the inserted disk is a CD and the desired format of the image to be generated is an ISO file ImgBurn will create a BIN file and a CDRWin file. If the inserted disk is a DVD instead an IMG file will be generated, in addition to a DVD and/or Media Descriptor Image layout files, only if these two files have been selected in the Settings section.

3.5.2 Test Results

Nine discs, all without copy-protection schemes were read correctly and working ISO image files which could be rendered via emulation were generated. The remaining four discs, all including different protection schemes (TOC, ProtectCD VOB and SecuRom) proved unsuccessful in creating working image files.

The maximum reading speed provided by ImgBurn is 56x (8.624 MB/sec) for CDs, 56x (77.56 MB/sec) for DVDs and 56x (252 MB/sec) for Blu-ray disks. During the image generation process the maximum reading speed recorded however was 12x (1.848 Mb/sec) for a CD and 12x (16.62 MB/sec) for a DVD.

4. SUMMARY AND CONCLUSIONS

Media transfer presents a number of technical challenges when dealing with complex objects such as computer games. However, tools are available and developers have a vested interest in ensuring their applications are accessible even to casual users.

The results of the performed tests on the magnetic media transfer tools have shown that the complexity of the process increases sharply and that a relatively high degree of expertise is required to get the best out of the tools. Certain output formats such as FDI have been found to be more capable of handling non-standard or protected disks. Catweasel supports a wide range of media, but during the test its performance was relatively poor, as less than half the tested items were effectively captured. In contrast, the free application NibTools performed better, even if the source data processed, as previously mentioned, was technically less accurate. Disk2FDI has demonstrated to be the most robust of the three, trading off a significant increase in transfer time for a level of detail and accuracy that far surpasses the other tools. The question in this context perhaps is where preservationists would rather invest their time: how much time they would take to capture an image, or how long they may potentially have to spend re-capturing an image using multiple means. Overall, the performance gap between free tools and commercial ones in these tests was minimal. Furthermore, the open source nature of NibTools implies a surrounding community which could provide in the future testing and optimization for this tool, in addition to open access APIs for more technical users, thus enabling this application to be integrated potentially within any framework. For the purposes of KEEP, this is naturally highly advantageous. For memory organizations the capture of magnetic media probably requires investment in a more stable and powerful transfer tool such as Disk2FDI.

Regarding the optical transfer tools tested it can be said that all of these were relatively easy to set-up and to operate. No significant difference in terms of disk reading time was found among these transfer tools. ImageBurn has proven to be reasonably effective with older Windows and PC formats and it can be considered a good starting point. Of the commercial tools, Blindwrite offers perhaps the best combination of robustness and flexibility – its ability to manage console game discs is highly advantageous compared to the others (though it should be noted that tests were not carried out with such discs). Certainly, Clone CD and Alcohol120% have a similar level of robustness and also offer some copyright protection scheme circumvention, probably best suited for older discs.

As far as optical media is concerned, the biggest technical issue is handling copy protection schemes. This might be a serious issue for those attempting to preserve with complex commercial objects such as games. In these cases, a viable solution would be non-technical: close collaboration with industry, lobbying for the importance of preservation and the collection of not just disc images but source code. For example, in 1997, id Software released the source code of the classic game DOOM freely, along with pre-release alpha and beta versions of the game. This brought direct advantages to the company: they were able to use community ports (Boom to MBF to prBOOM)[22] as the basis of the source code for a new mobile version. This is a case study that preservationists can use in the argument for the advantages of releasing of older code. In addition to this, it makes the game a preservationist's dream. There is no question that preserving access to source code remains the primary goal of preservationists working with complex digital objects.

Our investigation has shown that even start-up preservation projects with limited resources can have access to reasonably robust tools - one of which has the distinct advantage of being open source - capable of basic media transfer operations. We recommend using BlindWrite for the transfer of information from optical media, and Disk2FDI for the transfer of information from magnetic media. All of these aforementioned tools should be used in conjunction with file characterization software such as

[22] http://www.idsoftware.com/iphone-doom-classic-progress

DROID[23] or JHOVE[24] in order to identify the produced image files.

In terms of integration within a single framework, access to APIs and source code of tools is a fundamental requirement if tools are to be packaged to a single application. KEEP's intention is to link distinct applications into a common framework that sits alongside existing tools, creating a supportive workflow management environment. Following from this initial study, we are now in a position to integrate the recommended tools within this environment for further testing, which will be reported in due course.

5. REFERENCES

[1] Lorie, R. 2001. Long term preservation of digital information. In *Proceedings of* the *1st ACM/IEEE-CS joint Conference on Digital libraries* (Roanoke, USA, June 24 - 28, 2001). JCDL '01. ACM, New York, NY, 346-352. DOI= http://portal.acm.org/citation.cfm?id=379726.

[2] Sinclair, P., Billenness, C., Duckworth, J., Farquhar, A., Humphreys, J. and Jardine, L. 2009. Are You Ready? Assessing Whether Organisations are Prepared for Digital Preservation. In *Proceedings of* the *Sixth International Conference Preservation od Digital Objects* (San Francisco, USA, October 5 - 6, 2009). iPRES '09. 174-181. DOI= http://escholarship.org/uc/item/8dd2m5qw.

[3] von Suchodoletz, D., Rechert, K. and van den Dobbelsteen, M. 2010. Software archives as a vital base for digital preservation strategies. In *Proceedings of the 5th International Conference on Open Repositories* (Madrid, Spain, July 6 - 9, 2010). DOI= http://en.scientificcommons.org/58572706.

[4] Anderson, D. 2011 [Forthcoming]. Layman's guide to the legal issues involved in software re-use and emulation. Deliverable D2.6 KEEP Project.

[23] http://droid.sourceforge.net

[24] http://hul.harvard.edu/jhove

Impact Assessment of Decision Criteria in Preservation Planning

Markus Hamm
Vienna University of Technology
Vienna, Austria
hamm@ifs.tuwien.ac.at

Christoph Becker
Vienna University of Technology
Vienna, Austria
becker@ifs.tuwien.ac.at

ABSTRACT

Significant progress has been made in clarifying the decision factors to consider when choosing preservation actions and the directives governing their deployment. The Planets preservation planning approach and the tool Plato have received considerable takeup and produce a growing body of knowledge on preservation decisions. However, experience sharing is currently complicated by the inherent lack of semantics in criteria specification and a lack of tool support. Furthermore, the impact of decision criteria and criteria sets on the overall planning decision is often hard to judge, and it is unclear what effect a change in the objective evidence underlying an evaluation would have on the final decisions.

This article presents a quantitative approach and tool to support the systematic assessment of criteria and their impact in preservation planning. We discuss the reconciliation of different quality models and present an analysis tool integrated with the planning tool Plato. We further apply our analysis method to a body of real-world case study material and discuss the results. The outcomes provide directions to optimise and automate decision-making, watch, and policy definitions at large scales, and to lower entry barriers by focussing on those aspects that have the strongest impact.

Categories and Subject Descriptors

H.1 [**Information Systems**]: Models and Principles; K.6.4 Management of computing and Information Systems; H.3 Information Storage and Retrieval H.3.7 Digital Libraries

Keywords

Digital Preservation, Decision Making, Multiple Criteria Decision Analysis, Preservation Planning, Utility Analysis

1. INTRODUCTION

Over the past years, considerable effort has been invested in analysing the factors contributing to decision making in digital preservation and the constraints posed by different scenarios, and in building decision making frameworks and tools. With current state-of-the-art procedures in digital preservation, we can create plans that treat a certain part of the content in a large repository. The planning tool Plato[1], created in the project PLANETS, has been applied to a number of real-world and pilot cases and is producing a growing body of knowledge [3, 9, 20].

Consider an identified preservation problem consisting of a set of digital material that is at risk of becoming obsolete. The material is held by an organization. There is a number of possible alternatives to resolve the identified issues, and a number of objectives and constraints that have to be considered. The preservation planning approach implemented in Plato presents a systematic method and tool to create a plan for this scenario. Decision makers represent goals and constraints in a hierarchy of objectives resolving into decision criteria. They evaluate alternatives against these criteria by applying controlled experimentation and automated measurements, and take an informed decision based on the resulting objective evidence. The finalized plan is fully documented, and it is fully traceable to the reasons underlying each decision. The planning tool provides guidance and automation in the planning procedure.

Despite this progress, however, a number of significant challenges remain and pose a substantial barrier towards the successful transition of the control of preservation operations from ad-hoc decisions towards continuous management. On the one hand, preservation planning in reality still is a rather isolated affair, where knowledge is only exchanged informally. Plans created in the planning tool Plato can be shared with others by making them public, and a number of these plans is available for analysis by a growing user community. However, until now there has been no systematic assessment of the impact of decision criteria. This is partly due to the fact that the specification of decision criteria used to be entirely based on individual scenarios. This implied a substantial variation in criteria definition until recently, when a standardized method of identifying, documenting and reusing criteria with defined semantics was introduced [2]. The automation in decision making processes is still limited by the fact that many information needs cannot be addressed automatically. Continuous management, however, requires systematic mechanisms and processes for information exchange and control.

The project SCAPE[2] is set to move forward the control of digital preservation operations from ad-hoc decision making to proactive, continuous preservation management, through

[1] http://www.ifs.tuwien.ac.at/dp/plato
[2] http://www.scape-project.eu/

a context-aware planning and monitoring cycle integrated with operational systems. This systematic improvement of decision automation requires an assessment of the criticality and the exact impact of decision criteria.

To provide this analysis, this article presents a method and tool support for the quantitative assessment of decision criteria in preservation planning. We build upon a significant body of work collected in the last years, which includes preservation plans for different types of content, models for preservation goals and criteria, and a basic taxonomy of categories which we base our analysis upon. We conduct an analysis of key factors and decision criteria considered in preservation decisions and their quantitative influence on evaluation and decisions.

The article is structured as follows. Section 2 describes related work in the areas of Multi-Criteria Decision Analysis, Preservation Planning and decision criteria for digital preservation decisions, and software quality models. Section 3 discusses the reconciliation of existing models. Section 4 discusses key issues in decision criteria analysis and impact assessment of criteria, while Section 5 shortly presents a decision factor analysis tool. Section 6 presents some results of applying the presented analysis approach to a growing body of knowledge created in real-world case studies. Finally, Section 7 discusses implications and presents an outlook on future work.

2. RELATED WORK

Preservation Planning is a key element of the OAIS model [12]. The upcoming ISO standard describing metrics for Repository Audit and Certification includes detailed requirements on planning procedures that have to be considered to achieve trustworthy decision making. These include, for example, the requirements to explicitly specifiy the *'...Content Information and the Information Properties that the repository will preserve'* [15]. Clearly, such a specification needs to build upon (1) a model for specifying such properties, (2) an assessment of the possible actions that the repository can employ to achieve its goals within the constraints posed by these properties, and (3) a method to evaluate whether the repository will be able to preserve these properties, in which form, and at which costs and risks. Models for specifying *transformation information properties*, as the OAIS calls them, or *significant properties*, as they are often referred to, have been discussed intensely over the last years [5, 8, 19]. The realistic evaluation of such properties requires objective evidence, repeatable measures, and thorough documentation. The Plato approach combines such an evaluation method and supports the automated and repeatable documentation of objective evidence through controlled experimentation and automated measurements. At its heart, the so-called *objective tree* specifies goals and objectives of a preservation scenario and breaks these aspects down into decision criteria that can be quantitatively determined. Figure 1 presents a simple illustrative example containing three *decision criteria* and one requirement node ('Correctness') that comprises two of the criteria.

Preservation planning is a typical case of multi-criteria decision analysis [6]. In taking preservation decisions, decision makers have to reconcile potentially conflicting and initially ill-defined goals and find the optimal solution within weakly defined organizational constraints. The approach followed in Plato builds upon a widely used approach that resolves

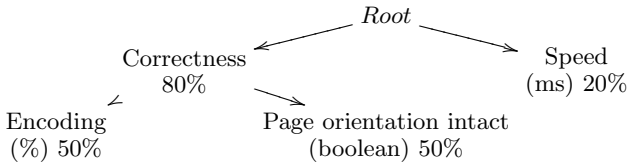


Figure 1: Highly simplified requirements tree

the incommensurability of multiple decision criteria by applying utility analysis [17]. To allow a comparison across the criteria, a *utility function* is specified for each criterion that contains an explicit mapping to a uniform utility score ranging from 0 (unacceptable) to 5 (best). This score can then be weighted and aggregated across the hierarchy.

The combination of objective evidence measured in specific scales, subjective assessment represented in case-specific utility functions, and relative weights across the goal hierarchy, is a powerful, yet flexible model. However, it requires a profound understanding of the intricacies of decision making scenarios, and a careful distinction between the key concepts of evidence, utility, and weighting [3]. Common approaches to sensitivity analysis vary the weightings of attributes to determine the robustness of assigned weights similar to the approach presented in [4].

The planning approach supported by Plato was also applied to bitstream preservation planning [23]. Recent discussions about preservation planning presented a categorization of decision criteria according to their measurement needs [2] and analysed a series of case studies, focusing on lessons learned and open challenges [3]. Kilbride discussed the fact that decision making can be very complex, and emphasized the benefits that experience sharing would provide for organizations facing the preservation planning problem [18]. McKinney compared Plato to a commercial implementation that follows a slightly simpler decision model [21].

One of the key aspects in planning is the question of software quality. The ISO standard 25010 - 'Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models' [16] is based on the earlier ISO 9126 family. The ISO/IEC 9126 standards [11] define a hierarchy of high-level quality attributes, where quality measures are based on procedures recommended in ISO 15939 [14]. SQUARE combines a revised quality model with evaluation procedures based on ISO 14598 [10]. It defines requirements on the specification of software product quality criteria [13]. Earlier, Franch proposed a six-step method for defining a hierarchy of quality attributes for a specific domain in a top-down fashion [7]. ISO 25010 states that it defines

- a *quality in use* model composed of five characteristics (some of which are further subdivided into subcharacteristics) that relate to the outcome of interaction when a product is used in a particular context. This system model is applicable to the complete human-computer system, including both computer systems in use and software products in use.
- a product quality model composed of eight characteristics (which are further subdivided into subcharacteristics) that relate to static properties of software and dynamic properties of the computer system. The model is applicable to both computer systems and software products.

Characteristic	ISO 25010 Definition
Functional suitability	degree to which a product or system provides functions that meet stated and implied needs when used under specified conditions, comprised of – Functional completeness: degree to which the set of functions covers all the specified tasks and user objectives – Functional correctness: degree to which a product or system provides the correct results with the needed degree of precision – Functional appropriateness: degree to which the functions facilitate the accomplishment of specified tasks and objectives
Performance efficiency	performance relative to the amount of resources used under stated conditions, comprised of – Time behaviour: degree to which the response and processing times and throughput rates of a product or system, when performing its functions, meet requirements – Resource utilization: degree to which the amounts and types of resources used by a product or system, when performing its functions, meet requirements – Capacity: degree to which the maximum limits of a product or system parameter meet requirements
Compatibility	degree to which a product, system or component can exchange information with other products, systems or components, and/or perform its required functions, while sharing the same hardware or software environment
Usability	degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use
Reliability	degree to which a system, product or component performs specified functions under specified conditions for a specified period of time
Maintainability	degree of effectiveness and efficiency with which a product or system can be modified by theintended maintainers
Portability	degree of effectiveness and efficiency with which a system, product or component can be transferred from one hardware, software or other operational or usage environment to another

Table 1: Software quality attributes as defined in ISO 25010 SQUARE [16]

Quality attributes are defined in a hierarchic manner. The quality model divides product quality into characteristics, each of which is composed of several sub-characteristics. Table 1 defines several characteristics relevant to preservation actions. Section 3 will discuss the relation of these to decision criteria in preservation planning.

These hierarchical structuring procedures have already been used to inform the hierarchical definition of objective trees in the planning approach in Planets. But since preservation planning has a specific focus, different compared to generic cases of software product selection [2], it is necessary to customize the quantitative part of evaluation, as recommended by ISO SQUARE.

Hence, the next section presents a quality model that is based on ISO 25010 for the high-level generic quality model and associates it with exemplary measurable criteria that have been of concern in productive decisions in preservation planning. This reconciled quality model then enables the analysis of accumulative decision factors such as the resource utilization of preservation action components in a systematic and standardized way, while retaining the full expressiveness and flexibility of the decision making framework.

3. RECONCILING DECISION MODELS

3.1 A generic taxonomy

A first in-depth analysis of about 600 decision criteria of planning studies led to a bottom-up classification of criteria according to their sources of measurement. This was discussed in detail in [2]. The primary distinction hereby is between criteria relating to a *preservation action* and criteria relating to its *outcome*. The latter is divided into *format properties*, *object properties* and *outcome effects* such as costs. This classification serves as a key tool to increase automated measurements in a measurement framework. However, it does not relate clearly to the impact that decision factors and criteria sets have on the final decisions for two reasons: (1) No impact analysis is performed, (2) Decision factors are related to concerns such as risks, which may be expressed by multiple criteria measured through diverse sources [2]. Thus, this article focuses on the top-down reconciliation of top-down models with the overall classification into *action* and *outcome* criteria. In particular, this section discusses format properties, software quality, and information properties.

3.2 Format Properties

The format website run by the Library of Congress (LoC) suggests to evaluate formats according to the two aspects *sustainability* and *quality and functionality*. Sustainability factors recommended are disclosure, adoption, transparency, self-documentation, external dependencies, impact of patents, and technical protection mechanisms [1].

PRONOM suggests to assess a given file format against each of the following characteristics and sub-characteristics:

- Capability: The support for features required or desirable to meet business requirements, such as support for specitic types of content (e.g. chart support in spreadsheet),
- Quality: The accuracy of information storage, represented by Precision and Lossiness.
- Resilience: Safety over time, represented by Ubiquity (resilience against obselescence), Stability (resilience against software updates), and Recoverability (resilience against accidental corruption).
- Flexibility: Ability to adapt to changing requirements, represented by Interoperability (with existing tools) and Implementability (the degree of difficulty to implement software for this format) [22].

The given list is not intended to be fully complete and needs customization and extension dependent on the given context. Furthermore, it is clear that most of these high-level factors are not directly measurable. While knowledge sources such as PRONOM document experts' assessments of some of these attributes, many characteristics are high-level characteristics and require assignment of more specific quantified properties to be reliably assessed. We use these factors for the high-level generic quality model and associate them with exemplary measurable criteria that have been of concern in productive decisions in preservation planning.

Figure 2 shows characteristics assembled from LoC and PRONOM (in bold letters) and links them to planning criteria extracted from several case studies. The characteristic 'impact of patents' was generalized into 'rights'. It can be seen that a combination of both models is required to cover all factors that have been used for evaluation in real-world decisions. Merging these references to a unified model as in the suggested model above leads to a more suitable model for the preservation context. Section 6 will shed some light on the actual impact that these format criteria have on real-world decisions in comparison to other decision factors such as preservation process requirements.

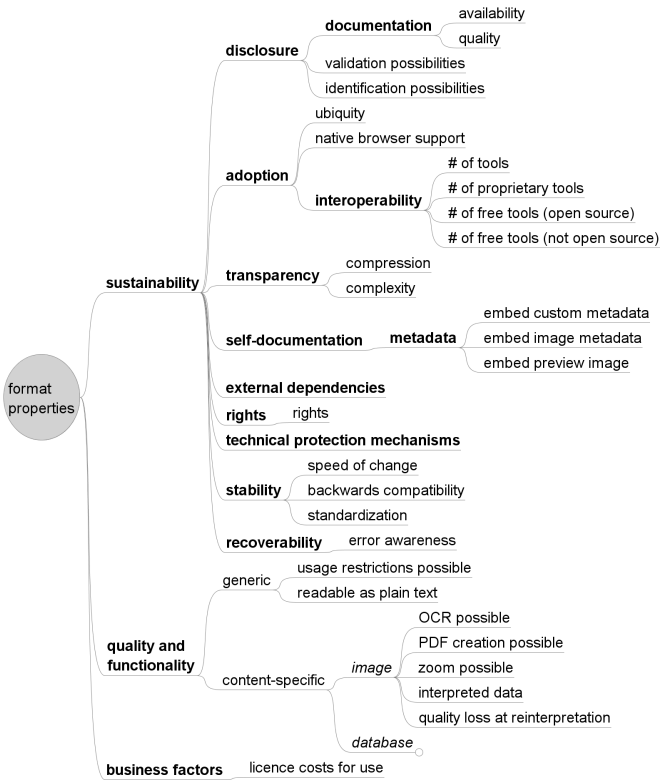


Figure 2: Format factors and associated criteria

On a more general perspective, the name *format* properties may be a bit misleading, since conceptually, this category can include any criterion referring to the *representation* of information in digital form, i.e. its encoding. This observation is particularly relevant in scenarios dealing with the preservation of large data sets instead of traditional 'file-based' objects.

3.3 Software Quality

The quality of software components has been analyzed extensively over the past decades, and a number of formal models have emerged. We analyzed decision criteria from planning case studies, based on previous analysis [2], and assigned them to the SQUARE quality model. Figure 3 illustrates a subset of criteria and their classification according to SQUARE. The ISO quality factors are given in bold.

The ISO quality characteristic *functionality* merits special attention. Functional *completeness* includes process-related features of software components such as the traceability of performed actions or the presence of mechanisms to support validation of input objects. However, content-specific features describing support of preservation action components for specific features of content also belong to this category. Functional *appropriateness* generally refers to the question whether certain preservation action components are applicable to an organization's holdings. This is generally not an evaluation criterion in planning, but rather a pre-selection criterion for creating the list of candidate actions that are evaluated. Finally and most crucially, functional *correctness* is at the heart of the quest for authenticity and represented as a specific category in the planning framework, as discussed below.

3.4 Information Properties and Functionality

The ISO characteristic *functional correctness* has an especially high relevance in the digital preservation context. Assuring that preservation action results are correct is a fundamental goal of digital preservation. This is covered by the category *Outcome Object* in the decision criteria taxonomy of Plato. Essentially, this can be further divided into

1. *Transformation Information Properties* refer to the significant properties to be preserved throughout changes of either environments or object representations.

2. *Representation Instance Properties* describe aspects of the representation, i.e. of the encoding, of information objects. This includes the file size required to represent a certain information object or the question if a representation is well-formed, valid and conforming to a certain expected format profile.

3. *Information Properties* are desired properties or features of the objects themselves.

3.5 Observations

The exact way of taking measures on criteria, measures which describe in a quantitative way the fulfilment of quality attributes, is a complex issue and highly domain dependent. The decision criteria taxonomy discussed in [2] provides important information about this and enables an additional classification that can be used to guide evaluation. More specifically, this means that some attributes can be researched, documented and fed into a catalogue; some are highly or entirely context-dependent, yet, they are relevant for selection and decision making; and some require empirical measures in controlled experimentation.

However, the taxonomy is not very meaningful with respect to criteria semantics. Hence, this section aimed at reconciling standard quality models with decision criteria. In particular, the ISO 25010 quality model presents an international standard for modelling software quality attributes in a high-level top-down fashion. This stable standard provides a solid reference to resolve ambiguities about the meaning of certain quality attributes such as reliability, stability, etc.

Clearly, the models discussed in this paper are all hierarchical. ISO has a hierarchical structure; the objective trees are hierarchical; the taxonomy of Plato is hierarchical. However, this does not mean that the quality model is an objective tree, or that the objective tree needs to conform to such a structure. There are many ways to structure hierarchical trees of criteria; the objective tree should contain all objectives and requirements that pertain to a certain scenario. The quality model *informs the definition* of such an objective tree. Similarly, the differentiation of the taxonomy described in [2] is essentially orthogonal to the ISO quality model. The taxonomy describes measurable criteria, not the concerns they relate to – it is a bottom-up classification, whereas the ISO model is a top-down quality model. For example, the ISO quality attribute 'performance efficiency' includes dynamic runtime criteria such as time used per sample object, but also static action criteria such as the capacity of a tool, e.g. the maximum number of files in a batch process. Thus, the models presented are complementary, and a combination of them is required to model the factors that have to be considered. This unification of models in concrete decision making is achieved within the planning framework.

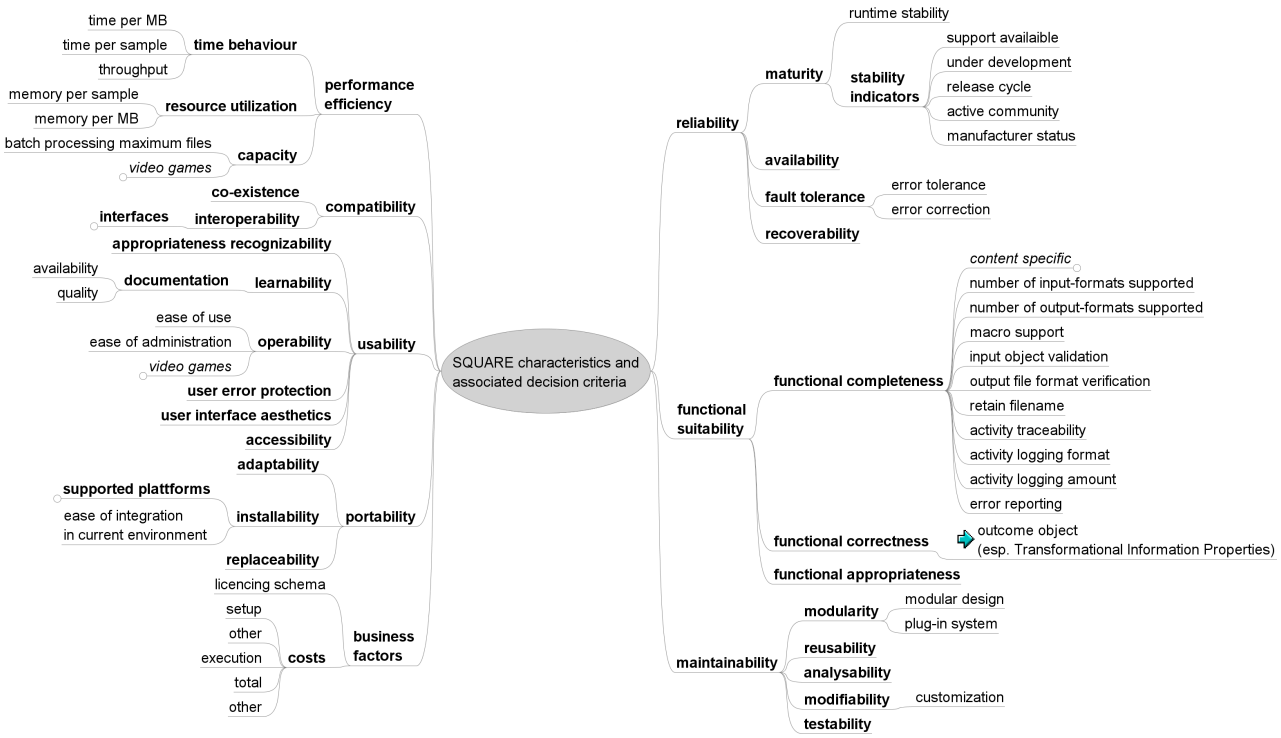


Figure 3: Metrics for SQUARE quality attributes

4. DECISION FACTORS ANALYSIS

As decision makers, we want to improve the efficiency of a specific decision making scenario while keeping full trustworthiness. For improving preservation planning processes in general, we want to improve efficiency over many scenarios. To advance the understanding of the field, finally, we want to gain insight into decision making processes and their key factors. We thus need to consider both single decision criteria as well as certain logical groupings of criteria. For a given set of decision criteria and plans, we want to answer several key questions:

1. What is the *impact* of a certain criterion on the decision? Would a change in its evaluation, i.e. in the objective evidence, change preference rankings on alternative solutions?

2. Considering a specific case: How critical was this criterion in other cases? Has it led to a rejection of potential alternatives in similar cases?

3. What is the *accumulated* impact of a set of criteria on decisions in certain scenarios? (For example, what is the accumulated impact of criteria relating to format risks in the preservation of scanned images in large libraries? What is the accumulated impact of the resource utilization of action components in large migration decisions in archives?)

4. Are there any sets of decision factors that are *dominated*, i.e. factors that by themselves cannot change decisions, no matter which evaluation values we insert?

5. What is the minimum set of criteria that have to be considered in a given scenario?

The questions relating to impact of a single criterion correspond to a robustness or sensitivity assessment. The previous approach to assessing sensitivity of decision makers' preferences computed variations of relative weightings to produce a robustness assessment judging the influence of tree branches on the root score. This does not address the specific scales, in particular the differences between numerical and ordinal measures. It also does not assess the sensitivity of the utility functions, which may include non-linear effects produced by the mappings. Furthermore, it does not consider reliability of measures [2]. The combination of these aspects, however, can lead to substantial variations in the scores, as we will see below. On the other hand, the questions regarding decision cases require an accumulated assessment of the impact of multiple criteria over sets of plans, where each criterion may appear in a number of plans. To achieve this, we will define impact factors for sets of criteria.

To answer the questions posed above, we need quantitative measures that consider

- the usage frequency and weight of a criterion in comparable scenarios (where a scenario is defined at least by the type of content and the type of organization), and

- the impact caused by a change in objective facts, i.e. the extent to which the utility scores of decisions including the criterion change when the evaluation facts change.

This requires us to integrate a number of properties in our assessment: (1) the number of times and frequency a criterion is used in planning cases, (2) the set of total weights of a criterion in each case, (3) the set of values collected for a criterion, and (4) the set of utility functions for the criterion.

ID	Factor	Definition
IF1	Count	Number of plans using this criterion
IF2	Spread	Percentage of plans using this criterion
IF3	Weight	Average total weight of this criterion
IF4	Discounted Weight	Sum of total weights of this criterion, divided by number of all plans
IF5	Potential	Average potential output range of this criterion
IF6	Range	Average actual output range of this criterion
IF7	Discounted Potential	Sum of all criterion potential output ranges, divided by number of all plans
IF8	Discounted Range	Sum of all criterion actual output ranges, divided by number of all plans
IF9	Maximum Potential	Maximum potential output range
IF10	Maximum Range	Maximum actual output range
IF11	Variation	Average relative output range
IF12	Maximum Variation	Maximum relative output range
IF13	Rejection Potential Count	Number of utility functions with an output range including 0.
IF14	Rejection Potential Rate	Percentage of utility functions with an output range including 0.
IF15	Rejection Count	Number of utility functions actually rejecting alternatives.
IF16	Rejection Rate	Percentage of utility functions actually rejecting alternatives.
IF17	Reject Count	Number of rejected alternatives.
IF18	Reject Rate	Percentage of rejected alternatives.

Table 2: Impact factors for single criteria.

In search for realistic, relevant and representative quantitative measures, we will define a number of impact factors for single criteria and groups of criteria. Section 6 will discuss the results obtained by their application to a set of of real-world results.

To consider the impact of criteria contained in a hierarchical structure, we have to consider their aggregation throughout the hierarchy. Criteria are weighted on all levels of the hierarchy in a relative fashion. To aggregate utility scores in the objective tree, the two standard weighted aggregation functions weighted sum and weighted multiplication are included in Plato. For weighted multiplication, utility values are taken to the power of the weight of the node to ensure that nodes with a weight of 0 result in a neutral element. The *total weight* of a criterion can be easily determined by multiplying its weight with all parent weights up to the root node of the tree.

Table 2 summarizes and names all impact factors, designated *IF*, for single criteria. The basic impact factors of a criterion are the number of plans referring to it, the average total weight of the criterion across these plans, and the relation between these. Let $C = \{c_1, c_2, ...c_n\}$ be the set of criteria and $P = \{p_1, p_2, ...p_m\}$ be the set of plans considered – for example, all plans that refer to the preservation of images in a library setting. Then for a criterion $c \in C$, P_c is the set of plans using c. Thus our first impact factor IF1 represents the size of P_c: $IF1(c, P) = |P_c|$. Let thus *IF1* be the *number of plans using criterion c* and *IF2* the *percentage of plans using criterion c*, i.e. $IF2(c, P) = \frac{|P_c|}{|P|}$. Let further be *IF3* the *average total weight of c* in plans where it is used as given in Equation 1, and *IF4* the sum of total weights divided by the size of the entire set P. IF4 thus includes a discounting for criteria that are rarely used, but with high average total weights.

$$IF3(c, P) = \frac{\sum_{i=1}^{k} w_{c,p_i}}{|P_c|}, p_i \in P_c \quad (1)$$

These simple factors do not represent the actual *impact* that a change in evaluation has, since they do not account for the utility function. Arguably, this utility has more impact on the final result than the weighting itself [2]. More meaningful impact factors of a decision criterion can thus be quantified by considering the possible effect that a change in the objective facts that the criterion refers to has on the assessment of the criterion with respect to the decisions taken. This can be obtained by calculating the change in the final score of the objective tree *root* caused by a change in the criterion evaluation. Consider a boolean criterion c with $values = \{Yes, No\}$. Let the utility function u defined in a certain plan p map *Yes* to a target utility of 5 and *No* to the target utility 1, i.e. $u_{c,p}(Yes) = 5, u_{c,p}(No) = 1$. If c is assigned a total weight $w_{c,p}$ of 0.25 in the given plan, the *potential output range por(c,p)* of criterion c in plan p is given by the weighted difference between the highest and the lowest possible utility result. Hence, in our case it is $(5 - 1) \times 0.25 = 1$. If $c \notin p$, the output range for (c, p) is considered 0. The theoretic maximum of all output ranges here is determined by the range of the utility scale, which in the case of Plato ranges from 0 to 5. In addition, $\sum_{i=1}^{k} por(c_i, p) \leq 5.0, c_i \in p$.

However, in fact no value v_c, p in this plan may actually be *No*. Thus, the *actual output range aor(c,p)* of criterion c in plan p is given by the weighted difference between the highest and the lowest result of the utility function applied to the actual evaluation values $v_c \in p$, as given informally in Equation 2, with $aor(c, p) \leq por(c, p) \forall c \in C, p \in P$. Similar calculations can be made for numeric criteria, for which thresholds define the utility function.

$$aor(c, p) = w_{c,p} \times (\max(u_{c,p}(v_{c,p})) - \min(u_{c,p}(v_{c,p}))) \quad (2)$$

Decision criteria often are defined defensively, i.e. potential bad outcomes are considered despite the fact that they are unlikely to happen. To investigate how likely potential bad outcomes actually are for certain criteria and candidates, we are thus interested in the ratio between *potential* and *actual* impact. This *relative* output range (or *Variation*) $ror(c, p) = \frac{aor(c,p)}{por(c,p)}$ corresponds to the question how far output ranges are in reality represented in the evaluation values or whether the occurring variance is much lower than the expected possible output range of a criterion.

Apart from the output ranges averaged over all plans using a criterion, we can also relate the sums of potential and actual output ranges to the total number of plans to account for the frequency of usage. This is in particular relevant if we are not looking at a scenario and a criterion, but rather analyzing a set of scenarios and criteria.

Finally, a discrete, non-weighted aspect has to be considered. If a utility function contains the target 0 in the output, it has the potential to reject an alternative as unacceptable, independently of the criterion weight. This is a crucial element of the decision method [3]. We are thus interested in (a) the *rejection potential* of a criterion, i.e. the utility functons with an output range including 0, (b) the *rejection*

ID	Factor	Definition
SIF1	Spread	Average spread of the criteria in the set
SIF2	Coverage	Percentage of plans using at least one of the criteria
SIF3	Weight	Sum of discounted average total weights
SIF4	Potential	Sum of discounted average potential ranges
SIF5	Maximum potential	Maximum compound potential ranges
SIF6	Range	Sum of discounted average ranges
SIF7	Maximum range	Maximum compound actual ranges
SIF8	Variation	Average of the relative output ranges
SIF9	Maximum variation	Average maximum of the relative output ranges
SIF10	Rejection Potential Count	Number of utility functions with output range including 0.
SIF11	Rejection Potential Rate	Percentage of utility functions with output range including 0.
SIF12	Rejection Count	Number of utility functions rejecting alternatives
SIF13	Rejection Rate	Percentage of utility functions rejecting alternatives
SIF14	Reject Spread	Percentage of plans affected by a reject out of this set
SIF15	Reject Count	Number of alternatives rejected.
SIF16	Reject Rate	Percentage of alternatives rejected.

Table 3: Impact factors for *sets* of criteria

of a criterion, i.e. the amount of utility functions that reject alternatives due to a utility of 0, and (c) the *rejects* of a criterion, i.e. the amount of alternatives rejected.

When analyzing criteria *sets*, we need slightly adapted impact factors. While factors such as count and spread can be aggregated in a straightforward way, others would lead to misleading figures. For instance, simply summing up the average weights would neglect the fact that these averages are calculated based on the partial set P_c. To analyze criteria sets over the entire set P, we can thus only sum up *discounted* average weights. Table 3 lists the resulting impact factors for criteria sets.

While this set of factors is mathematically simple and robust, it is clearly somewhat redundant. However, the exact factor to be used for answering a certain question has to consider a number of dimensions. To reduce the set of factors that need to be analyzed to answer specific questions and provide guidance on concrete analysis tasks, Section 6 will present analysis results for all factors on a set of 210 criteria from six case studies selected in a homogeneous problem space.

5. TOOL SUPPORT

To support the systematic and repeatable assessment of decision criteria, we are developing an interactive, web-based analysis tool. This tool is compatible with the planning tool Plato and can be seen as a complementary addition to the primary planning workflow. It will thus enable decision makers to share their experience and in turn leverage the wisdom of their community's peers in anonymized ways by aggregating the experience that planners wish to share.

The tool loads preservation plans from the planning tool's knowledge base (provided the plan has been released by

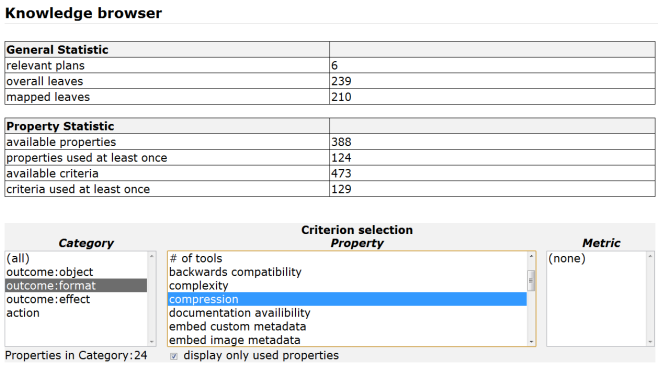


Figure 4: Knowledge browser criteria navigation

the owner and approved by a moderator). It processes and anonymizes plans and presents the decision maker or analyst with a number of features that facilitate systematic analysis in search of answers to the questions posed above:

- The planner can select a set of plans to be considered, i.e. filter the scenario set to be analyzed.
- The planner can then dynamically select properties of interest. For each property, the tool calculates all impact factors described.
- The tool furthermore visualizes several attributes of interest for each property, such as the different utility functions defined in various plans, in graphical form.
- Finally, to enable the analysis of not single criteria, but *criteria sets*, the user can dynamically create hierarchical property sets that reflect natural groupings of criteria such as all format properties that are considered relevant. The user can thus analyze the properties of aggregate sets of criteria in flexible configurations. We will discuss several such sets in the next section.

Figure 4 shows a screenshot crop of the navigation part, where the user can browse categories and properties of interest. Upon selection of a property or its associated metric, the tool visualizes a number of analysis results. The next section discusses these in detail.

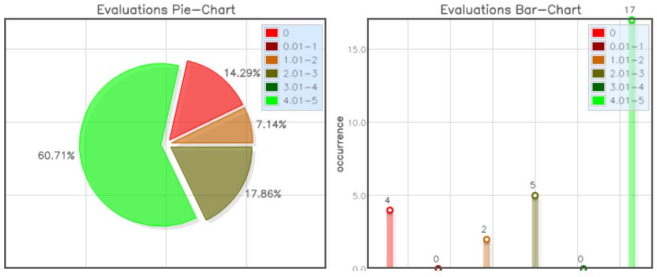
6. ANALYSIS OF RESULTS

To illustrate the application of the above calculations and investigate the usefulness of our method and tool to answer the questions posed in the beginning, we analyse a set of related real-world case studies. Our analysis case includes the 6 plans shown in Table 4, which is a subset of the plans outlined in [3], where all plans deal with image preservation. They contain a total of 239 decision criteria, of which 210 (87,9%) have been mapped to uniquely identified properties. (The remaining decision criteria all occur only in one plan and have a Rejection Potential of 0.) Out of 473 criteria currently available in the knowledge base of the planning tool, 129 are of relevance in the analysis set.

The tool enables us to browse the criteria categories, select criteria, and analyze their properties and behaviour both in detail and through visualization. Figure 5 shows the tool displaying a visualization of the decision criterion *Format compression*. This is an *ordinal* criterion with the possible values *None, Lossless*, and *Lossy*. It is used frequently, in

	Organization type	Planning set	Criteria	Mapped	Alternatives	Chosen action
1	National Library	Large collection of scanned images in TIFF-5 (80TB)	24	24	7	Convert to JPEG 2000
2	National Library	Large collection of scanned images in TIFF-6 (72TB)	43	35	5	Keep status quo, see [20]
3	National Library	Collection of scanned high-resoluton images in TIFF-6	35	29	3	Keep status quo
4	National Library	Small collection of scanned images in GIF	26	25	4	Convert to TIFF-6 (ImageMagick)
5	Professional photographer	Digital camera raw files (CRW,CR2,NEF)	69	67	7	Convert to DNG (lossless) with Adobe DNG Converter
6	Regional archive	Digital camera raw files (NEF)	42	30	5	Convert to TIF (Photoshop CS4)

Table 4: Selected case studies on image preservation.



Ordinal Transformer	0.0	0.5	1.0	2.0	3.0	5.0
mapping	Lossy				None	Lossless
mapping	lossy					lossless, none
mapping		lossy		lossless		none
mapping	lossy					lossless, none
mapping			lossy		lossless	none

Figure 5: Visualization of *Format compression*

5 of 6 plans, with an average total weight of 0.0276. The average potential output range is only 0.13, but 60% of the utility functions have rejection potential. The top left pie chart shows a distribution over utility output ranges. We can see that almost 15% of values are rejected. The top right shows a frequency distribution along the utility scale. On the bottom, we see the anonymized utility functions defined by the five plans in which this property was used. Clearly, lossy compression is always the worst case, but only in 3 out of 5 cases is it a reason for immediate rejection of an alternative. *None* is the option with the highest scores on average, but in one case, considered worse than *Lossless* compression. The accumulated knowledge can also be used to gain insight about typical preferences and support proactive recommendation of utility settings. The fact that lossy compression is in all utility functions dominated by lossless and compression-free encoding comes as no surprise as it corresponds to common knowledge in the community. In other cases, it will be valuable input for a recommender function that can base recommended utility curves for certain users on the accumulated insight of others having tackled comparable problems. In the case of *lossless* vs. *none*, it can be seen that there is no dominating value, since the preference of lossless vs. lossy compression depends on a number of factors [3].

Figure 6 shows a raw view on the most frequently used criteria as displayed in the current version of the analysis tool. Clearly, the meaning of all these numbers is not immediately accessible to a decision maker and will require interpretation by systematic tools, since the question which factors to consider depends entirely on the scope of interest. Essentially, non-discounted factors will be of interest once we have decided to include a decision criterion or criteria set: They refer to the set of plans that use the criterion or set. On the other hand, if we have not decided upon inclusion or are not thinking about a concrete scenario, we need discounted factors to investigate the relative importance and the cumulative impact across multiple plans. Similarly, the pure counts are not very helpful and the corresponding indicators only become meaningful when used relatively with respect to the size of the criteria set and the size of the set of plans. However, indicators such as the *rejection potential* of criteria can provide good indicators for the criticality of a certain aspect of interest.

The raw statistics of single criteria thus present an important basis on which to assess specific criteria in certain situations. However, for the purpose of this paper, logical criteria sets such as those discussed in Section 3 are much more interesting. To illustrate the accumulated impact of such sets, we used the property hierarchy builder in the analysis tool and specified a number of criteria sets in correspondence to the models discussed above. Figure 7 shows these sets and their impact factors. While space constrains a full analysis and discussion, a number of observations can be drawn.

Format criteria are relevant in all plans, with a coverage of 100%. Their compound weight is 0.18. They achieve a maximum compound range of 0.86. On average, format properties exhaust a maximum of 33% of their utility range. The criteria set contains 17 utility functions with rejection potential. Every second plan in our set is affected by actual rejects caused by these criteria. Performance efficiency, on the other hand, has rejection potential, but none of the tested alternatives was rejected because of performance efficiency drawbacks.

Several aspects of actions are normally included in evaluation, but have very little impact on the decisions (Maintainability, Usability, Portability, Reliability). Business factors, which include costs and licensing, have a much higher relevance. *Representation Instance Properties*, such as *Format is well-formed and valid*, have a high rejection potential and do lead to rejection in one case.

The most important group of criteria, of course, is concerned with *significant properties* (Transformation Information Properties), which can also seen as belonging to the functional correctness of performed actions. Every third plan is affected by a reject caused by a loss of authenticity in content preservation actions. The maximum compound change caused by criteria of this set is substantial with 1.28. We can further see the impact factors of the specific subset of 12 criteria describing different metrics to assess image

Impact Factors Table																					
Category	Property	Metric	IF1	IF2	IF3	IF4	IF5	IF6	IF7	IF8	IF9	IF10	IF11	IF12	IF13	IF14	IF15	IF16	IF17	IF18
OUTCOME_FORMAT	documentation availibility		6	100%	0,018	0,018	0,0812	0,0417	0,0812	0,0417	0,1562	0,1562	0,3667	1	2	33,33%	1	16,67%	1	3,23%
OUTCOME_FORMAT	standardization		6	100%	0,0238	0,0238	0,0976	0,0821	0,0976	0,0821	0,2	0,2	0,713	1	2	33,33%	1	16,67%	1	3,23%
OUTCOME_FORMAT	compression		5	83,33%	0,0276	0,023	0,1319	0,0815	0,1099	0,0679	0,2295	0,2	0,5133	1	3	60%	1	20%	4	14,29%
OUTCOME_OBJECT	image: similarity	equal	5	83,33%	0,0272	0,0227	0,1252	0,1002	0,1043	0,0835	0,3125	0,3125	0,6	1	2	40%	2	40%	2	8,33%
OUTCOME_OBJECT	relative filesize		5	83,33%	0,0273	0,0227	0,1362	0,0752	0,1135	0,0627	0,24	0,1562	0,5867	1	5	100%	1	20%	1	3,57%
OUTCOME_OBJECT	image: colour space	equal	4	66,67%	0,0303	0,0202	0,1497	0,0781	0,0998	0,0521	0,3125	0,3125	0,25	1	3	75%	1	25%	1	5,26%
ACTION	interoperability: batch processing support		4	66,67%	0,0189	0,0126	0,0799	0,0212	0,0533	0,0142	0,11	0,085	0,25	1	1	25%	0	0%	0	0%
ACTION	performance efficiency: time per MB		4	66,67%	0,0638	0,0425	0,3188	0,1998	0,2125	0,1332	0,625	0,5	0,5988	0,8	4	100%	0	0%	0	0%
ACTION	portability: ease of integration in current environment		4	66,67%	0,0484	0,0323	0,2056	0,1403	0,1371	0,0935	0,5	0,5	0,4125	1	2	50%	0	0%	0	0%
OUTCOME_FORMAT	speed of change		4	66,67%	0,0237	0,0158	0,0929	0,0421	0,0619	0,0281	0,126	0,084	0,4583	0,6667	1	25%	0	0%	0	0%
OUTCOME_FORMAT	ubiquity		4	66,67%	0,036	0,024	0,168	0,1326	0,112	0,0884	0,35	0,28	0,6722	0,8889	1	25%	0	0%	0	0%
ACTION	usability: ease of use		4	66,67%	0,0281	0,0187	0,0961	0,0481	0,0641	0,032	0,16	0,16	0,375	1	0	0%	0	0%	0	0%
ACTION	business: costs: setup: software licence inital		3	50%	0,0322	0,0161	0,1608	0,0754	0,0804	0,0377	0,3	0,2096	0,2996	0,6988	3	100%	0	0%	0	0%
ACTION	business: licencing schema		3	50%	0,0229	0,0115	0,0983	0	0,0492	0	0,11	0	0	0	1	33,33%	0	0%	0	0%
ACTION	functional suitability: activity traceability		3	50%	0,03	0,015	0,1358	0,0817	0,0679	0,0408	0,17	0,17	0,5333	1	2	66,67%	0	0%	0	0%
ACTION	functional suitability: output-file format verification		3	50%	0,0382	0,0191	0,1263	0	0,0631	0	0,17	0	0	0	0	0%	0	0%	0	0%
OUTCOME_OBJECT	image: height	equal	3	50%	0,0258	0,0129	0,0938	0	0,0469	0	0,18	0	0	0	2	66,67%	0	0%	0	0%
OUTCOME_OBJECT	image: width	equal	3	50%	0,0258	0,0129	0,0938	0	0,0469	0	0,18	0	0	0	2	66,67%	0	0%	0	0%
OUTCOME_FORMAT	# of free tools (not open source)		2	33,33%	0,0075	0,0025	0,0362	0,0312	0,0121	0,0104	0,0625	0,0625	0,5	1	1	50%	1	50%	1	14,29%
OUTCOME_FORMAT	# of free tools (open source)		2	33,33%	0,0106	0,0035	0,0519	0,0519	0,0173	0,0173	0,0938	0,0938	1	1	1	50%	1	50%	1	14,29%
OUTCOME_EFFECT	archival storage		2	33,33%	0,2425	0,0808	0,79	0,665	0,2633	0,2217	1,08	1,08	0,75	1	0	0%	0	0%	0	0%
OUTCOME_EFFECT	automated quality assurance possible		2	33,33%	0,0538	0,0179	0,2688	0,2	0,0896	0,0667	0,4	0,4	0,5	1	2	100%	1	50%	4	33,33%
ACTION	business: costs: execution: software develop new features		2	33,33%	0,0189	0,0063	0,0944	0,0354	0,0315	0,0118	0,1063	0,0495	0,4	0,6	2	100%	0	0%	0	0%
ACTION	business: costs: execution: software licence running		2	33,33%	0,0191	0,0064	0,0956	0	0,0319	0	0,1063	0	0	0	2	100%	0	0%	0	0%
ACTION	business: costs: setup: initial hardware		2	33,33%	0,0875	0,0292	0,4375	0,05	0,1458	0,0167	0,625	0,1	0,2	0,4	2	100%	0	0%	0	0%
ACTION	business: costs: setup: other		2	33,33%	0,035	0,0117	0,15	0,1	0,05	0,0333	0,2	0,2	0,5	1	1	50%	0	0%	0	0%
ACTION	business: costs: setup: staff external expertise needed		2	33,33%	0,0356	0,0119	0,1781	0	0,0594	0	0,25	0	0	0	2	100%	0	0%	0	0%
ACTION	business: costs: total		2	33,33%	0,0513	0,0171	0,2563	0,205	0,0854	0,0683	0,3125	0,25	0,8	0,8	2	100%	0	0%	0	0%
OUTCOME_FORMAT	embed custom metadata		2	33,33%	0,0232	0,0078	0,0994	0,042	0,0331	0,014	0,1148	0,084	0,5	1	0	0%	0	0%	0	0%
ACTION	functional suitability: error reporting		2	33,33%	0,0361	0,012	0,1705	0,09	0,0568	0,03	0,18	0,18	0,5	1	1	50%	0	0%	0	0%
OUTCOME_OBJECT	image metadata: tiff properties	retained	2	33,33%	0,0102	0,0034	0,0239	0,0239	0,008	0,008	0,0412	0,0412	1	1	0	0%	0	0%	0	0%
OUTCOME_OBJECT	image: bits per sample	equal	2	33,33%	0,0108	0,0036	0,05	0	0,0167	0	0,0688	0	0	0	1	50%	0	0%	0	0%
OUTCOME_OBJECT	image: ICC colour profile	equal	2	33,33%	0,0108	0,0036	0,0431	0,0275	0,0144	0,0092	0,055	0,055	0,5	1	0	0%	0	0%	0	0%
OUTCOME_OBJECT	image: image quality		2	33,33%	0,0492	0,0164	0,2282	0,2282	0,0761	0,0761	0,3125	0,3125	1	1	1	50%	1	50%	1	9,09%
OUTCOME_OBJECT	image: resolution	equal	2	33,33%	0,0141	0,007	0,0562	0,0183	0,0281	0,0092	0,0825	0,055	0,3333	1	0	0%	0	0%	0	0%
OUTCOME_OBJECT	image: resolution unit	equal	2	33,33%	0,0244	0,0081	0,1063	0,1063	0,0354	0,0354	0,1575	0,1575	1	1	0	0%	0	0%	0	0%
OUTCOME_OBJECT	image: similarity	SSIM simple saturation	2	33,33%	0,0197	0,0066	0,0984	0,0357	0,0328	0,0119	0,15	0,0714	0,238	0,476	2	100%	0	0%	0	0%
OUTCOME_OBJECT	image: similarity	Equality judgement	2	33,33%	0,0197	0,0066	0,0811	0,06	0,027	0,02	0,12	0,12	0,5	1	0	0%	0	0%	0	0%
OUTCOME_OBJECT	image: similarity	SSIM simple hue	2	33,33%	0,0197	0,0066	0,0984	0,0236	0,0328	0,0079	0,15	0,0471	0,157	0,314	2	100%	0	0%	0	0%
OUTCOME_OBJECT	image: text quality		2	33,33%	0,0492	0,0164	0,2282	0,2282	0,0761	0,0761	0,3125	0,3125	1	1	1	50%	1	50%	1	9,09%
ACTION	portability: runs on Linux		2	33,33%	0,0042	0,0014	0,0158	0,0138	0,0053	0,0046	0,0276	0,0276	0,5	1	0	0%	0	0%	0	0%
ACTION	portability: runs on MacOS		2	33,33%	0,0041	0,0014	0,0153	0	0,0051	0	0,0276	0	0	0	0	0%	0	0%	0	0%
ACTION	portability: runs on Windows		2	33,33%	0,0041	0,0014	0,0165	0	0,0055	0	0,0276	0	0	0	0	0%	0	0%	0	0%

Figure 6: Impact factors of all frequently ($\geq 33\%$) used criteria, sorted by descending IF1 (count)

similarity (ranging from error measures and advanced metrics such as structural similarity to subjective assessment). It can be seen that they account for a significant part of the significant properties.

Finally, all 64 action criteria together (without functional correctness) have not a single reject. However, they achieve an accumulated potential output change of 2.5 and realize this potential to a large degree, with a maximum compound actual range of 1.25.

7. DISCUSSION AND OUTLOOK

This article has presented a systematic approach for analysis of decision criteria in preservation planning. We discussed the reconciliation of quality models describing properties of preservation action components as well as formats. We further presented a method and tool for quantitative impact assessment of decision criteria, and discussed results of applying the assessment to a series of real-world case studies.

One application of this analysis will be a reduction of complexity and manual effort of preservation planning through a reduction of the number of alternatives that are evaluated in depth. This can be supported by an early filtering of candidate actions based on correlating organizational goals and constraints (expressed in semantic models) with documented knowledge as well as experience shared by other organizations.

The outcome of this analysis thus provides directions to optimise and automate decision-making, watch, and policy definitions at large scales. It also enables smaller organizations to make decisions at a lower entry barrier by prioritizing essential evaluation factors. For example, criteria that have no rejection potential and a very low potential output range could be left out in preliminary evaluation or substituted with aggregated experience from analogous scenarios. While this cannot in all cases replace a full-depth evaluation, it can provide a tool for trade-off decisions between risks and costs and serve as a basis for solid planning. Similarly, it can increase the focus and impact of research in characterization and Quality Assurance by prioritizing aspects that have the strongest impact.

The identification and unambigous specification of decision criteria across scenarios and organizations has additional benefits. In addition to the questions posed in the beginning of this paper, questions such as 'Is a given set of criteria complete for the scenario at hand? Which other criteria may be relevant?' are of relevance and need to be systematically addressed. This can for example lead to proactive recommendation of decision factors to stakeholders, and to an intelligent monitoring service that raises alerts when certain conditions have been discovered to be of relevance by stakeholders outside an organization. This accumulated experience sharing shall advance the knowledge base of DP operations considerably and enable the transition to a continuous and continuously optimizing management activity.

Current and future work is focused on expanding the data set to cover additional scenarios such as those described in [3, 9], completing the assignment of metrics to the SQUARE quality model, and using the information obtained to provide advanced, proactive decision support within the preservation planning procedure. This includes a further quantitative in-depth assessment of impact factors to arrive at a reduced and simple-to-use set of factors targeted at specific investigation scenarios.

Acknowledgements

Part of this work was supported by the European Union in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

Name	size	SIF1	SIF2	SIF3	SIF4	SIF5	SIF6	SIF7	SIF8	SIF9	SIF10	SIF11	SIF12	SIF13	SIF14	SIF15	SIF16
Format	31	25,27%	100%	0,183	0,812	1,396	0,435	0,864	0,327	0,42	17	36,17%	6	12,77%	50%	8	25,81%
Action: Performance Efficiency	7	11,9%	83,33%	0,048	0,234	0,625	0,155	0,5	0,228	0,257	4	80%	0	0%	0%	0	0%
Action: Functional Completeness	15	13,33%	83,33%	0,063	0,261	0,428	0,115	0,244	0,239	0,303	5	41,67%	0	0%	0%	0	0%
Action: Maintainability	3	5,56%	16,67%	0,003	0,013	0,08	0,003	0,02	0,083	0,083	0	0%	0	0%	0%	0	0%
Action: Usability	6	11,11%	66,67%	0,019	0,064	0,16	0,032	0,16	0,062	0,167	0	0%	0	0%	0%	0	0%
Action: Portability	5	33,33%	100%	0,036	0,153	0,5	0,098	0,5	0,182	0,4	2	20%	0	0%	0%	0	0%
Action: Reliability	8	4,17%	33,33%	0,009	0,035	0,129	0,007	0,04	0,062	0,062	0	0%	0	0%	0%	0	0%
Action: Business factors	16	20,83%	83,33%	0,124	0,601	1,335	0,195	0,366	0,187	0,269	17	85%	0	0%	0%	0	0%
Action: All	64	15,1%	100%	0,314	1,415	2,502	0,619	1,25	0,164	0,241	29	50%	0	0%	0%	0	0%
Representation Instance Criteria	12	18,06%	100%	0,053	0,236	0,734	0,063	0,156	0,049	0,083	5	38,46%	1	7,69%	16,67%	1	3,23%
Information Criteria	57	1,17%	33,33%	0,033	0,152	0,625	0,152	0,625	0,035	0,035	2	50%	2	50%	16,67%	1	9,09%
Transformation Information Criteria	80	16,88%	100%	0,188	0,817	1,285	0,363	0,876	0,58	0,62	16	19,51%	3	3,66%	33,33%	2	6,45%
Image Similarity Criteria	12	16,67%	83,33%	0,047	0,222	0,69	0,13	0,401	0,148	0,256	7	58,33%	2	16,67%	33,33%	2	8,33%
Outcome Effects	3	27,78%	50%	0,109	0,395	1,48	0,309	1,48	0,583	0,833	2	40%	1	20%	16,67%	4	26,67%
Outcome Object: All	149	11,07%	100%	0,274	1,205	1,409	0,578	1,406	0,335	0,36	23	23%	6	6%	33,33%	3	9,68%

Figure 7: Criteria sets and their cumulative impact factors as shown in the analysis tool

8. REFERENCES

[1] C. R. Arms and C. Fleischhauer. *The Digital Formats Web site*. The Library of Congress, accessed September 2011. `http://www.digitalpreservation.gov/formats/`.

[2] C. Becker and A. Rauber. Decision criteria in digital preservation: What to measure and how. *Journal of the American Society for Information Science and Technology (JASIST)*, 62(6):1009–1028, June 2011.

[3] C. Becker and A. Rauber. Preservation decisions: Terms and Conditions Apply. Challenges, Misperceptions and Lessons Learned in Preservation Planning. In *Proc. JCDL 2011*, June 2011.

[4] J. Butler, J. Jia, and J. Dyer. Simulation techniques for the sensitivity analysis of multi-criteria decision models. *European Journal of Operational Research*, 103(3):531 – 546, 1997.

[5] A. Dappert. Deal with conflict, capture the relationship: The case of digital object properties. In *Proceedings of the 7th International Conference on Preservation of Digital Objects (iPRES2010)*, Vienna, Austria, September 2010.

[6] J. Figueira, S. Greco, and M. Ehrgott. *Multiple criteria decision analysis: state of the art surveys*. Springer Verlag, 2005.

[7] X. Franch and J. Carvallo. Using quality models in software package selection. *IEEE Software*, 20(1):34–41, Jan/Feb 2003.

[8] S. Grace, G. Knight, and L. Montague. *InSPECT Final Report*. InSPECT (Investigating the Significant Properties of Electronic Content over Time), December 2009. `http://www.significantproperties.org.uk/inspect-finalreport.pdf`.

[9] 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), June 2010.

[10] ISO. *Information technology – Software product evaluation – Part 1: General overview (ISO/IEC 14598-1:1999)*. International Standards Organization, 1999.

[11] ISO. *Software Engineering – Product Quality – Part 1: Quality Model (ISO/IEC 9126-1)*. International Standards Organization, 2001.

[12] ISO. *Open archival information system – Reference model (ISO 14721:2003)*. International Standards Organization, 2003.

[13] ISO. *Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Measurement reference model and guide (ISO/IEC 25020:2007)*. International Standards Organisation, 2007.

[14] ISO. *Systems and software engineering – Measurement process (ISO/IEC 15939:2007)*. International Standards Organisation, 2007.

[15] ISO. *Space data and information transfer systems - Audit and certification of trustworthy digital repositories (ISO/DIS 16363). Standard in development*, 2010.

[16] ISO/IEC. *Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models (ISO/IEC 25010)*. International Standards Organisation, 2011.

[17] R. L. Keeney and H. Raiffa. *Decisions with multiple objectives: preferences and value tradeoffs*. Cambridge University Press, 1993.

[18] W. Kilbride. Preservation planning on a spin cycle. *DPC What's New*, 28, 2010.

[19] G. Knight and M. Pennock. Data without meaning: Establishing the significant properties of digital research. *International Journal of Digital Curation*, 4(1):159–174, 2009.

[20] 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 Magazine*, 15(11/12), November/December 2009. `http://dlib.org/dlib/november09/kulovits/11kulovits.html`.

[21] P. McKinney. Preservation planning: A comparison between two implementations. In *7th International Conference on Preservation of Digital Objects (iPRES2010)*, Vienna, Austria, September 19-24 2010.

[22] The National Archives of the UK. Evaluating your file formats. `http://www.nationalarchives.gov.uk/documents/information-management/evaluating-file-formats.pdf`, accessed September 2011.

[23] E. Zierau, U. B. Kejser, and H. Kulovits. Evaluation of bit preservation strategies. In *7th International Conference on Preservation of Digital Objects (iPRES2010)*, Vienna, Austria, September 19-24 2010.

Simulating the Effect of Preservation Actions on Repository Evolution

Christian Weihs, Andreas Rauber
Vienna University of Technology
Vienna, Austria
{weihs,rauber}@ifs.tuwien.ac.at
http://www.ifs.tuwien.ac.at/dp

ABSTRACT

One of the most important challenges in planning and maintaining a digital repository is to predict the needed resources on a long term basis, especially storage size and processing power. The main problem emerges from the need to migrate the data at certain times to newer file types, which takes time and alters the needed storage space, potentially branching into several migration paths for individual objects. Understanding the effect of different policy decisions, such as when to migrate or whether to stay within a format family or branching into several format families turns into a complex task, specifically when considering non-trivial ingest structures and assumptions on format evaluations. In this paper we present *ReproSim*, a framework that simulates the evolution of a digital repository and helps predicting these factors. We demonstrate the complexity and power of simulation to assist in preservation decisions in a set of scenarios involving different ingest and preservation planning profiles.

1. INTRODUCTION

Designing and operating a digital repository is a complex task. Especially estimating the scaling of the repository system, i.e. estimating the required storage space and computational power, across time considering a range of environmental options poses non-trivial challenges. While assumptions about the number and expected size of new objects to be ingested can be made with some diligence, the need for preservation actions to keep digital objects accessible adds significantly to the complexity. Following a migration strategy as one feasable way to maintain objects accessible, objects are converted to (potentially several) new formats at certain intervals in time, where the frequency of such migrations usually will depend on the validity and accessibility of a specific format (family). Thus, after a certain number of iterations, each object may exist in several versions, branching into a tree of different format (families), each of which again will be subjected to subsequent migrations. Identifying the effect of certain migration policy decisions, i.e.

iPRES2011, Nov. 1–4, 2011, Singapore.

- when to migrate: at ingest? when a specific format version is due to loose support? two months after the next-plus-one generation of the format comes into being?
- in how many paths to migrate: just within the format family? convert into more stable alternative formats that require fewer subsequent migrations? combinations?
- which tool (complexity) to use for migration: e.g. computational requirements such as more resource-demanding better-quality tools vs. simpler tools for mass-migration, effects on storage efficiency of the resulting objects
- for which files to apply these strategies, depending on file size, ingest type,..

is a complex issue. Identifying when peaks in computational resources for mass migrations are to be expected, or how storage requirements will grow, and how these change as a consequence of more risk-averse or risk-taking preservation policies requires detailed simulation of a repositories behavior based on explicitly modelled assumptions and specifications. This allows to understand the effect of certain policy decisions, specifically with respect to the branching factor of migrations into several target formats, providing a better basis to understand the trade-off between less risk (several copies in different formats) vs. more focused strategies.

Given the complex dependencies of such format decisions, bundled with non-linear growth both of the number as well as the size of objects to be ingested results in repository configurations that make straighforward calculation of its evolution unfeasible. Simulation offers a powerful approach to better understand the characteristics of a repository as it evolves under certain assumptions. Specific scenarios can be modelled and compared against each other, the effect of different policies can be analyzed, with subsequent decisions being based on the result of clearly specified simulation parameters rather than mere estimates. These, in turn, allow a monitoring of the validity of the simulation, as the actual evolution of crucial parameters such as ingest volumes, format validity periods, as well as the computational and storage costs of specific actions are tracked. This provides solid guidance in managing complex repository systems dealing with large volumes of heterogeneous material that are to be preserved over time.

To facilitate this process we have developed a simulator, that can show how a certain repository configuration will look like after several decades under a range of conditions that can be specified flexibly as a set of simulation parameters. It supports the specification of a repository configuration (file types, sizes and ingest timestamps/frequency, as well as future evolution of these) based on configuration files, as well as based on an existing collection profiles. Different migration rules can be specified, and the effect of these subsequently verified when running the simulation. The state of the archive at each point in time in terms of computational resources, storage requirements, and the number of versions of each object as well as of entire subcollections can be evaluated from the resulting data structures and logs.

The remainder of this paper is structured as follows: Section 2 reviews related work on preservation planning and collection profiling, forming the basis of simulating a repository's evolution. Section 3 describes the architecture and simulation parameters for the repository simulator. Exemplary simulation runs are presented in Section 4, followed by a short summary and outlook on future work in Section 5.

2. RELATED WORK

While being commonly used to understand the behavior of complex systems, modeling and simulation do not have a strong history in the analysis of digital repository systems. One of the few systems simulating aspects of a repository is ArchSim [7]. Focusing on storage technology, it allows to simulate the mean time to failure of an archive based on a library of failure distributions. The probability of not being able to interpret a format as it ages and becomes obsolete, for example, is modeled by a Weibull distribution. Different failure models can be assumed for different storage technologies. Using a complex architecture of triggers allows efficient modelling of failure probabilities over long periods of time. ArchSim/C [8] explicitly models costs associated with operating archival storage, including costs for creating, operating, monitoring and repairing a complex storage system. In a related line of work, a modeling approach is presented in [6] to analyze the reliability of system configurations for digital preservation. Again, the focus is on understanding the effect of component failures within a storage system. While these studies focus on understanding system failure characteristics and associated costs, the simulator presented in this paper focuses on understanding the evolution of individual files across a series of migrations into multiple branches, and the associated requirements in terms of storage and computational resources.

Testing and evaluating the effect of preservation action has been more intensively addressed from a planning perspective [2, 4]. Specifying the requirements for a specific preservation challenge (also referred to as objectives) and measuring how well different tools perform on selected sample data provide solid evidence for decisions on which preservation action component to deploy, and that component's effect on the object (in terms of significant properties retained), the storage space required, as well the the complexity of the deployment with respect to system and human resources that need to be provided. Preservation planning thus, on the one hand, provides valuable input to the simulator, concerning information on the processing requirements of certain types of preservation action tools as well as the resulting changes in object storage size. These can be obtained either directly from measurements obtained in preservation planning [3] or dedicated benchmark experiments measuring tool performance and the effect of preservation actions in controlled settings [1]. On the other hand, the repository simulator presented in this paper provides valuable input to a preservation planning process by providing a basis to estimate the costs associated with a certain preservation action, thus effectively closing the loop between planning and evaluation in these criteria.

In addition to the model parameters specified for a simulation run, more realistic initial configuration can be obtained from collection profiling services, as well as format registries such as PRONOM [5], which provide consolidated information on the lifetime/support time for selected formats, as well as offering a basis for analyzing the evolution of formats.

3. SIMULATING REPOSITORY EVOLUTION

The goal of the repository simulator is to offer the possibility to specify the content and ingest behaviour of a repository, then simulate migration rules on the files in the repository based on preservation plans, and collect statistical information about the changes in the repository.

During the simulation basically three different types of events get processed:

- **Ingest of new objects:** New files are added to the repository. The characteristics of this input stream, specifically date, initial file type and object size can be configured.
- **Migration of a file:** A file needs to be migrated to one or several other file types. The moment a migration has to take place can be specified by a set of rules depending on a range of factors, for example the expiration date of the file type or the size of the file.
- **Collect statistics:** On a regular basis statistical information is collected and stored in a file for later evaluation. This includes average file size, the number of executed migrations and so on.

3.1 Architecture

The Repository Simulator is realized as a Java application. The simulated archive is stored in a MySQL database, accessed via Hibernate. The following objects are mapped into the database model:

- **StoredFile:** Each ingested object is stored as a file stub (i.e. a file's profile) in the database. If a file gets migrated, a new file instance's profile is stored to the archive.
- **FileType:** Describes the available file types. Every file type consists of a type family name (ie. "Word Document") and a subtype name, which specifies the exact type version (ie. "Word 6.0"), as well as the average date of validity and the periodicity with which new versions are released.

- **MigrationTool:** The main properties of a migration tool are the duration of the process and how the file size is altered during the migration.
- **MigrationRule:** A migration rule defines when and how a migration should take place. This includes conditions which need to be met for the rule to be triggered (e.g a rule should only trigger for files smaller/larger than a certain size), the scheduled moment of the migration, the destination type (or a list of types, if for example a word document should be transformed into a PDF and a plain text file) and the used tools.

The model provides all information on the processes in the repository. On each migration event the new file is linked with its direct ancestor and with the tool element of the migration path. Additionally a generation counter gives direct insight how often a certain file got migrated. Furthermore, each file carries the information when it has been generated, by which tool and which rule. This makes it easy to track down the complete history of a certain file and allows to analyze the number of versions present of each original file or group of files, the percentage of active (ie. leaf versions in the migration tree) or inactive (ie. files that have already been migrated to newer versions) as well as the storage space used by these. (Note that this distinction between active and inactive files is currently rather basic. More complex representations may need to be modeled to account for the fact that an object may be migrated to a different format family to mitigate potential risks by having two different active versions in two format families. Extensions such as these are currently being implemented as part of the first evaluation cycle of the system. Similarly, delete orperations are currently being added to account for specific delete operations on migration, e.g. always keeping the original but deleting intermediate versions.)

3.2 Configuration

The configuration of a repository is done in plain text *ini*-files. There are basically four types of configuration files needed to specify all aspects of the simulation. In the following the core configuration possibilities are listed. To make the configuration flexible, many fields in the configuration are parsed with an expression language library, which means that any mathematical term can be used. Those fields are: the quantity and file size in the ingest configuration, the term and condition fields of the migration rule, and resulting change in file size and the computational cost of the migration process for each tool, expressed either in computation time or e.g. processor cycles (both of which can subsequently be normalized across time as the computational power of the underlying hardware infrastructure evolves). Note that file sizes in the configuration have no special magnitude (bytes, kilobytes etc.) associated with them. They can be specified in any magnitude, as long as it is in all configuration files.

- **Repository:** In this file the start and the end of the simulation is configured, as well as (the sequence of) all ingest events. It basically describes the characteristics of the repository to be simulated.

  The base configuration can either be obtained by a collection profile from an existing archive, by providing a list of file profiles, or by specifying groups of objects and their ingest characteristics by listing the number of objects, the mean and standard deviation in file size. This will create the according set of objects following e.g. a Gaussian or Weibull distribution. Similarly, the timeline of the ingest process can be modelled, so not all files are ingested at the beginning of the simulation, but the repository can grow step by step via ingest of original objects (in addition to the migrated ones) during the simulation process. in the case of modelling an existing repository for simuation purposes, the state of the repository needs to be provided as a collection profile detailing either the individual objects and their characteristics (format, sizes, ingest timestamps), or as a more compressed representation creating a model of the repository. Furthermore, parameters allow the specification of the growth characteristics of an archive, both in terms of number of objects and the average file size. These parameters can either be estimated or taken from the current history of a repository, e.g. the increase in average filesize of a collection of digital photographs or powerpoint files across the years, as well as the increase in numbers. These can be specified via almost arbitrary complexity, ranging from simple linear growth to more complex functions fitting real-life growth curves. Additionally, in combination with the format family configuration described in more detail below, the ingested objects will be of a specific version of the given file type families according to the timstamp within the simulation progress. Starting the simulation then creates the respective "files" as simulated entities with the respective ingest timestamps in the database.

  Listing 1 shows a sample repository configuration. In this case the simulation runs over 40 years starting in 2010/01/01 and ending 2049/12/31. In the beginning 1000 files of the type "doc" are inserted. They have an average size of 10000 with a deviation of 500. The next files are added in the year 2020: 3000 jpg files with average size 25000, deviated by 700. This ingest group has also specified the attributes "successive interval" (with value 2) and "successive count" (value 10). This way it is possible to specify a repetitive ingest, in this case every 2 years the same ingest is repeated for a total of 10 times.

  To make the configuration easier there is a simple collection profiling tool included that takes the repository structure from an existing archive, allowing one to operate on a real-life object type distribution. This currently reads objects from a mounted files system and can be adapted to meet API requirements of specific repositories. Alternatively, a statistics report that may be exportable from a repository can be converted to match the configuration file. Currently, the evolution parameters have to be estimated from an existing collection profile manually, with plans to provide this as an integrated module being currently evaluated.

Listing 1: Repository Configuration

```
[repository]
 simulation_start=2011/01/01
```

```
simulation_end=2050/12/31

[ingest1]
 Type=doc
 quantity=1000
 filesize = Dist:normal(10000,500)
 ingest_date=2011/01/01

[ingest2]
 Type=jpg
 quantity=3000
 filesize = Dist:normal(25000,700)
 ingest_date=2020/01/01
 successive_intervall=2
 successive_count=10
```

- **Filetype:** This describes a file type family. Each file type family consists of a family name and several subtypes with a specified time frame of how long they are supported, as well as how frequently new subtypes are generated. This results in a set of available file types at each point during the simulation, with objects being ingested as new originals usually being created in the most recent version of a file format family available. More complex configurations of mixes during overlap periods of format version validity are in principle possible using a number of distribution functions such as Gaussians or Weibull distributions.

  Listing 2 provides the configuration for the format type family "video". A subtype "xvid01" is created which is valid from the year 2000 to 2011. To model a sequence of consecutive subtypes a repetition group with the last two properties in the example can be specified. In this case 50 subtypes are created, shifted by 5 years, so the second one is valid from 2005 to 2017. Note that these subtypes can be specified at different levels of granularity, either, as shown in the example below, simple in the form of "AVI" files, or at a more detailed level, representing a range of video codecs embedded in an AVI container as individual subformats. This allow a realistic recreating of repository settings. Again, in principle the specifications of format version validities need to be specified in the simulation model. These settings could, in principle, also be imported from format registries.

Listing 2: file type configuration

```
[FileType]
name=video
extension=.avi
type=video

[subtype1]
subtype=xvid01
created=2000
expired=2012
successive_intervall=5
successive_count=50
```

- **Migration rule:** The migration rule is the most important part of the configuration. Each rule has an effective date, a source type for which the rule should be triggered, and a condition (for example "current file size is smaller than 5000") that determines for which files the rule should be executed. This allows for a rather fine-grained specification of migration policies, e.g. migrating smaller objects to multiple formats, whereas very large files might be migrated only within a single format family strand. Furthermore, migrations can either be based always on the most recent format version of each object, or always be based on the originally ingested objec, i.e. the root object, by setting the "source" parameter of the migration rule.

  Beyond that one or more destination types along with the tools to be used for the simulated migration can be listed. For each destination type one can again specify a condition, so complex migration policies can be mapped into the simulation model.

  Listing 3 provides a rule to migrate all files of the type family "doc" to the format families "pdf" and/or "rtf". The property "term" describes when the rule has to trigger, in this case two months before the respective sub-version of the file format expires. Two destination types are specified. The first one is a type of the family "pdf". The subtype is not specified directly, but with the keyword "maximal step" it is indicated that we want to migrate to a type from that format family which is available at migration time and has the longest expiration time. The destination subtype for rtf is specified as "minimal step", which means that the subtype with the next higher expiration date should be taken.

  Both destinations have a condition specified. The migration to each destination is only executed if it evaluates to true. In this example files smaller than 15000 are migrated only to rtf and files larger than 10000 only to pdf. Files with a size between 10000 and 15000 are migrated to both destination formats. (This example is only supposed to demonstrate the flexibility of configurations, allowing to address space considerations that may appear in real preservation planning scenarios, when certain objects types that may be in demand by different user communities should be made available in different formats, whereas other, potentially very large files, should not be kept in duplicate versions. It is not supposed to represent a recommended preservation plan within the scope of this paper. The same applies to the timing settings provided, i.e. whether a migration should happen 2 months prior to the expiry date.)

Listing 3: migration rule configuration

```
[migrationrule]
description=migration for doc files
term=subtype_expired -2*month
source_Type=doc

[destinationformat1]
destination_Type=pdf
destination_SubType=[maximal_step]
condition=file_size > 10000
tool=doc2pdf
source=current
```

```
[destinationformat2]
destination_Type=rtf
destination_SubType=[minimal_step]
condition=file_size < 15000
tool=doc2rtf
source=root
```

- **Migration tool:** For each virtual migration tool one can specify how the size of the file is changed during the migration process and how long the migration will take. Both the file size and migration duration are specified using mathematical expressions. Thus, one is not bound to simple linear changes but more complex effects can be simulated. Both units are dimensionless, i.e. as for th especification of the file sizes, these can be given in Bytes, Kilobytes, etc. More importantly, for the simulation of computational resources, either computation time or e.g. processor cycles may be specified. The latter may proove useful when more realistic estimates of the computational requirements are required. By mapping operation cycles in a virtual unit, normalization factors may be applied to account for improvements in processing power over time. Still, in first experiments, specifying actual processing time, and then applying a normalization factor to account for improved computational facilities, seemed to be more easily accepted. Note, that the primary use of the effort simulation is not a precise determination of the HW requirements at a specific point in time in the future, but to capture the potential of cumulative effects resulting from certain preservation policies. These may stem, for example, from the difference of migrating on-ingest (usually leading to a more even spread of subsequent migrations) or on-expiry - resulting in strong peaks if all objects of a specific format version need to be migrated to, e.g. the subsequent version.

  Two examples:

  - **size=currentsize * Math:log(currentsize)**: This specifies a logarithmic growth of the file. (The keyword "Math" in this string references the Java class java.lang.Math, which has methods for many mathematical operations, all accessible through this keyword.)
  - **duration=Math:max(currentsize * 3, 18000)**: The migration should take three times as long as the file is big, but at least 18000ms.

From this set of configurations the simulation of a repository's evolution is started. For each (set of) files specified in the repository configuration, the according sets of files are "created" as database entries with the respective timestamps. For each of these the respective migrations based on the preservation planning triggers as specified in the migration rules setting are executed consecutively. Thus, for each file specification in the database meeting a migration condition the respective new file(s) are generated with new ingest timestamps and file sizes considering the migration time needed and the file size change incurred as specified in the respective migration tool specification. From these, the resulting hypothetical computational load (i.e. the number of files to be migrated at any specific point in time) and the required storage space for the accumulated archive can be calculated. (Currently, the simulator only supports single-processor migration, i.e. all pending migrations are executed consecutively. An extension allowing the specification of a (growing) multiprocessor architecture or simulated cloud support to scale with the repository is under investigation).

4. EVALUATION

For evaluation we performed a series of simulations to see whether the simulation works correctly, and to what extent the available configurations are flexible enough to support realistic scenarios. In the following one of our sample configurations is presented along with the results of the simulation.

4.1 Configuring a repository

In this simulation, we defined 6 format families. Note that the types may not conform to reality, because they are just an assumption to provide a simple simulation example - in principle, these could also be modeled as entirely abstract format types that have certain characteristics such as stability, support, etc. that are relevant for the aspects covered by the simulator. The same applies to the migration tools, which can be specified in an equally abstract way. We decided to choose real file types for the sake of clarity and ease of discussion. Please note, again, that the validity periods specified do not correspond to real values, which would need to be obtained from format registries or from an analysis of the evolution of object formats in an existing repository. The same applies to the file sizes. The types are the following:

- **jpg:** for compressed pictures; every subtype is valid for 28 years and every 15 years a new subtype becomes available.
- **tiff:** for uncompressed pictures; valid for 44 years with a new subtype every 30 years.
- **wordDoc:** for proprietary text documents; valid for 15 years, new subtype every 5 years.
- **openDoc:** for editable text in a free format; valid for 8 years and every 4 years a new subtype.
- **pdf:** valid for 16 years and a new subtype every 8 years.
- **rtf:** valid for 12 years and a new subtype every 6 years.

The simulation runs for 100 years and every year there are 100 wordDoc files with average size of 500 and 100 jpg files with average size of 1600 are added to the repository.

Several rules specify the migrations:

- **jpg:** The compressed pictures are migrated to the tiff format. The file size multiplies by 10 during the migration.

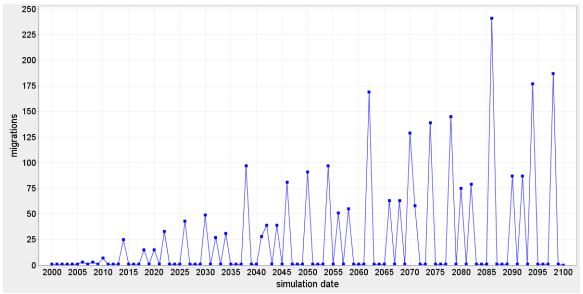


Figure 1: Number of migrations to be expected at each time interval

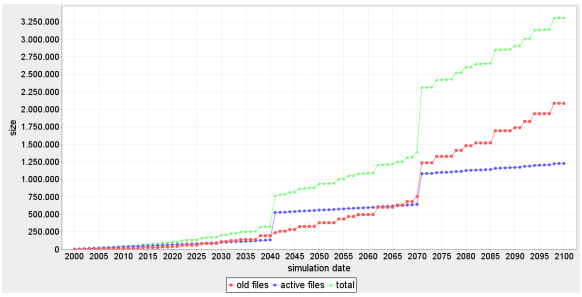


Figure 2: Number of files in the repository, subdivided in active files as well as earlier interim migration copies

- **wordDoc:** Word Documents get migrated both to openDoc, rtf and pdf. The file size grows 20 percent in case of openDoc and decreases by 5 percent to rtf and 10 percent to pdf.
- **pdf, tiff, rtf and openDoc:** These types get migrated to a newer version in the same family several months before they expire. All are migrated with maximal step size and the file size varies several percent downwards in case of tiff, upwards for the other ones.

For all migrations tools are created with the size changes as described above and named following the pattern "<source>2<destination>" (eg. "jpg2tiff").

4.2 Simulationg migrations

Figure 1 shows the number of migrations performed as the archive grows and format versions trigger migration. There are very few migrations in the first years of the simulation, but the number increases as time passes and more format versions expire. There are several peaks that show the moments multiple types expire at the same time. These peaks are a good hint that the preservation plan may be revised to avoid them or to plan for appropriate resources for mass migration projects at regular intervals. It also allows a more detailed evaluation of slight modifications of certain rules, e.g. starting migrations at earlier points in time, migrating always to the most recent version, potentially suffering from lower-quality tools available vs. migrating to a more stable version that already exists for a longer period of time, and others.

Figure 2 shows the storage space needed by the repository. The total size is growing constantly, with several boosts correlating partially to the peaks in Figure 1. Note that the two very big steps in the years 41 and 71 have no special spike in Figure 1. These two steps are the result of the expiring of a sub-version of the tiff format. We do not have especially many tiff files in the repository, but the files are significantly larger than the other files, so the migration of those files has a drastical impact on thel archive size assuming that the original files are not deleted.

At the same time, the proportion between the active files (representing the most recently migrated version of each file in each migration branch) and the old files (the copies left behind after migrating a file to a newer subtype) changes drastically. This provides a good impression whether a cleanup

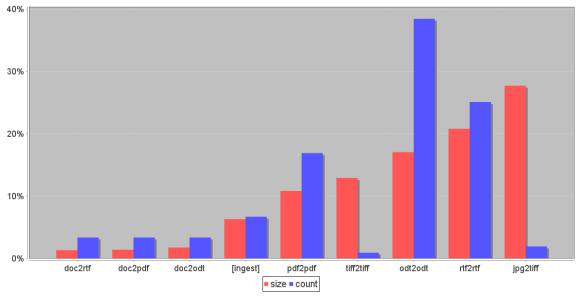


Figure 3: Tool usage, depicted as data volume handled by a tool and number of times it was called

strategy for the old data may be necessary to save on storage costs, potentially adapting preservation policies at an institution. (Different delete policies for inactive file formats, such as deleting the last-but-one, every second version, or keeping only the original and the most recent version, etc. are currently being evaluated as an additional configuration setting.)

In Figure 3 the usage of the tools is analyzed. The first bar for each tool family shows the percentage of the aggregated size of all files generated with this tool. The second bar shows how often this tool-family is called. In the diagram it is easy to see that the tools "odt2odt", "jpg2pdf" and "rtf2rtf" have the biggest potential to save space as they are responsible for 71% of the whole archive content. But one should keep in mind that "odt2odt" and "rtf2rtf" are called many times. In comparison "jpg2tiff" is called infrequently, so it might be ok to exchange it with a tool that is slower but has a smaller output.

Beside the statistics generated by the simulator, it is also of interest whether the simulation process is finished in reasonable time. For this simulation 259400 migrations were executed and the total process needed less than 15 minutes on a simple workstation, showing the feasibility to run sufficiently complex simulations.

4.3 Comparing policies

In the following example we assume an archive ingesting files from two format families. For ease of discussion, let's call them documents and images. We further assume that two different image subformats exist (e.g. jpeg and tiff), one

with a rather rapid release cycle of 3 years, the other with a slower cycle of 7 years, whereas for the document file formats we assume a single format family with a replacement cycle of 5 years. The individual format versions receive between 13 to 25 years of support, defining at each point in time a number of potential migration versions. To keep the graphs simple, ingests are kept growing linearly for all formats, with annual ingests. For preservation, jpegs are migrated both to jpeg and tiff, whereas both tiff and the documents are migrated within their respective formats. The two different policy strategies, and thus the only parameters varied in this simulation, concern the step-width for the migrations, i.e. whether we prefer to migrate each object to its next available version (min-step) or to the newest version available at the time of migration (max-step).

The results of this scenario are shown in Figure 4. The Min-step scenario obviously requires many more migrations, as multiple versions of each file are generated at rather short cycles whenever a format version expires, with the subsequent version expiring soon after. This results in increasingly high peak loads, especially in years when two format versions happen to expire at the same time. It also leads to a much higher number of "old" files, i.e. interim migration versions that soon surpass the number of objects actively used (note the different scales in the two graphs), calling for the evaluation of suitable deletion strategies, which are currently being added to the simulator.

5. CONCLUSIONS

Planning and operating a repository tasked with preserving heterogeneous sets of objects over long periods of time poses severe challenges when it comes to estimating growth in storage space as well as processing power required to run the required preservation actions. Also, from a preservation planning perspective, the effects of certain policy decisions such as when to migrate and how many copies to retain, are difficult to certain due to the complex behaviour emerging from multiply branching migration paths. Two different policies are devised, both relying on migration at ingest. In one case, objects are only migrated within the tiff family using the maxstep, i.e. migrating to the most recent format version available.

In this paper we presented *ReproSim*, a tool to simulate repository evolution over time. the strength of the approach lies in the flexibility offered by the configuration of the system, both in terms of modelling the content as received from the producers over time, as well as the content produced as a result of preservation actions, specifically migrations. Based on explicitly modelled assumptions of format stability, changes in object size induced by migrations, and others that can be explicitly specified (or, preferably, should be modeled based on an analysis of the history to date for the respective format types), different scenarios can be evaluated and compared, and the effect of different policies can be demonstrated. This, in turn, provides a better basis for policy, design, and structural planning decisions, helping both in the set-up and operation of a repository, as well as supporting preservation planning to evaluate the effect of certain recommendations.

While the simulator offers a very flexible basis for configuring different institutional settings, scenarios for object format evolution as well as preservation plans, a range of additional parameter settings emerged as being desirable from first case studies. Some of these, addressed already in the paper, include the possibility to simulate different object deletion policies, support for simulating multi-processor/cloud environments when evaluating peak loads, and others. A tighter integration with existing format registries, as well as more sophisticated collection profiling will allow better estimates for some of the core parameters in the system, specifically file format evolution and stability, as well as the characteristics of the ingest stream over time from different consumers. This will also allow verification of the simulation against existing repositories and their evolution to verify parameter settings and projections on file size growth. A tighter integration with preservation planning frameworks may help to provide closed feedback loops as well as offer better estimates on tool behavior with respect to resulting object sizes and processing times. Last, but not least, an improved interface helps with specifying the different scenarios and visualizing results in an integrated application.

6. REFERENCES

[1] Brian Aitken, Petra Helwig, Andrew Jackson, Andrew Lindley, Eleonora Nicchiarelli, and Seamus Ross. The planets testbed: Science for digital preservation. *Code4Lib*, (3), June 2008.

[2] Christoph Becker, Hannes Kulovits, Mark Guttenbrunner, Stephan Strodl, Andreas Rauber, and Hans Hofman. Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. *Intl Journal on Digital Libraries (IJDL)*, 10(4):133–157, Dec 2009.

[3] Christoph Becker, Hannes Kulovits, Michael Kraxner, Riccardo Gottardi, Andreas Rauber, and Randolph Welte. Adding quality-awareness to evaluate migration web-services and remote emulation for digital preservation. In *Proceedings of the 13th European Conference on Digital Libraries (ECDL 2009)*, volume 5714 of *LNCS*, pages 39–50. Springer, September 2009.

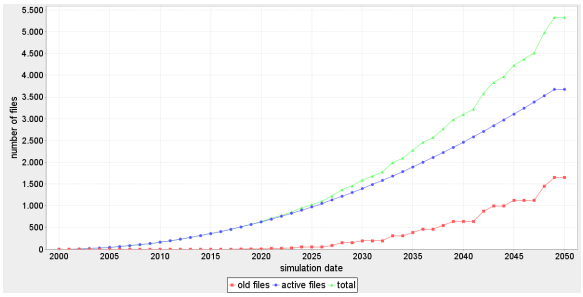
[4] Christoph Becker and Andreas Rauber. Decision criteria in digital preservation: What to measure and how. *Journal of the American Society for Information Science and Technology (JASIST)*, 62:1009–1028, 2011.

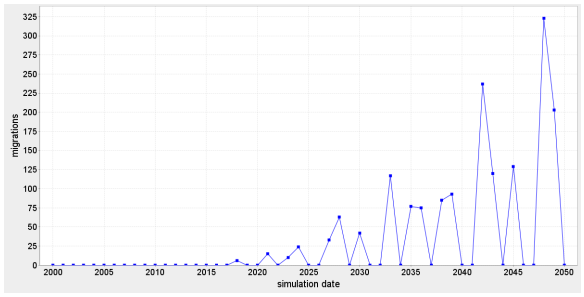
[5] Tim Brody, Leslie Carr, Jessie Hey, Adrian Brown, and Steve Hitchcock. PRONOM-ROAR: Adding format profiles to a repository registry to inform preservation services. *International Journal of Digital Curation*, 2(2), November 2007.

[6] Panos Constantopoulos, Martin Doerr, and Meropi Petraki. Reliability modelling for long term digital preservation. In *Proceedings of the 9th DELOS Network of Excellence Thematic Workshop on Digital Repositories: Interoperability and Common Services*, Heraklion, Greece, May 11-13 2005.

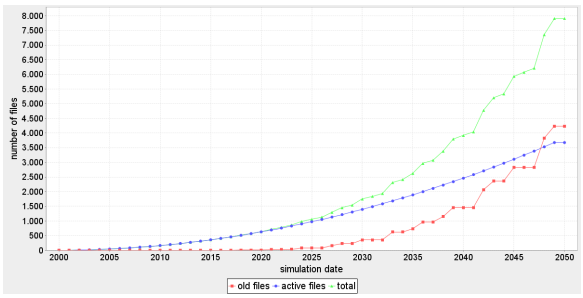
[7] Arturo Crespo. *Archival repositories for digital libraries*. PhD thesis, Stanford University, March 2003.

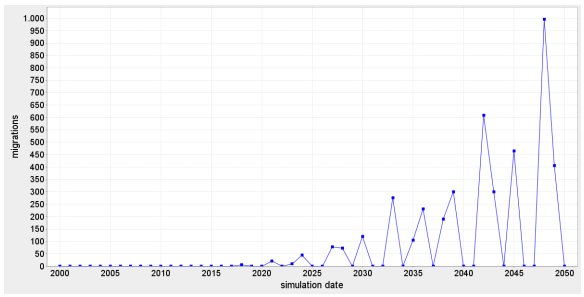
[8] Arturo Crespo and Hector Garcia-Molina. Cost-driven design for archival repositories. In *Proceedings of the First ACM/IEEE Joint Conference on Digital Libraries (JCDL'01*, pages 363–372, Roanoke, Virginia, USA, 2001. ACM Press.



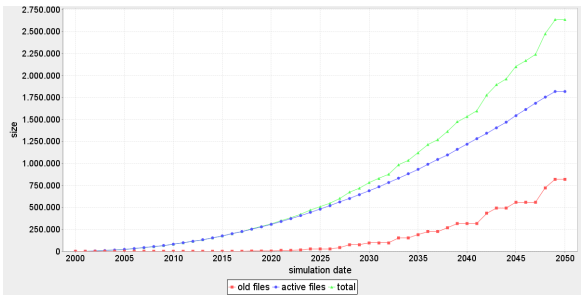


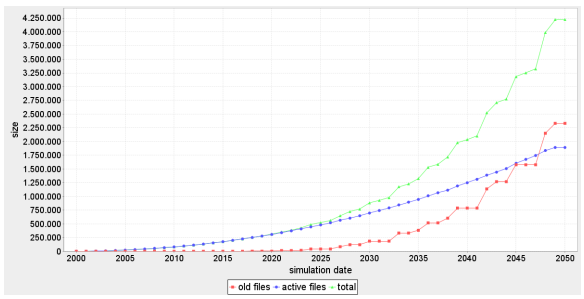
MaxStep: migrating to the newest available format version: (a) file count, (b) migrations





MinStep: migrating to the next available format version: (c) file count, (d) migrations





Resulting storage requirements: (c) MaxStep, (d) MinStep

Figure 4: Comparing migration step withs: Figs a and b show the number of files and migratons when migrating to the newest available file format version; Figs c and d show these for migrations to the respective next available versions. The resulting storage requirements and distributions between active and interim files is given in Figs (e) for MaxStep and (f) MinStep.

Risk Assessment in Digital Preservation of e-Science Data and Processes

Sara Canteiro
INESC-ID
Rua Alves Redol, 9
Lisbon, Portugal
s.canteiro@gmail.com

José Barateiro
INESC-ID, LNEC
Rua Alves Redol, 9
Lisbon, Portugal
jbarateiro@lnec.pt

ABSTRACT
Risk is a constant in every area and at all levels of any organization, whether in a general context or in a specific activity, project or function. Risk Management comprises a set of coordinated activities to direct and control an organization with regard to risk. Risk Assessment is considered the most important phase of Risk Management, which consists in identifying, analyzing and evaluating risks. Digital preservation's main concern is to keep information accessible and understandable over a long period of time, through means of digital objects; therefore, it is an area that needs a thorough Risk Management and, especially, a thorough Risk Assessment. In fact, the digital preservation process can be seen as Risk Management activities to protect digital information from inherent threats and vulnerabilities. The digital preservation problem can be even more complex in the context of e-Science, which is progressively being considered as a reference method for experimental scientific discovery, and whose data and processes need to be handled and preserved. As such, this paper analyzes the applicability of Risk Assessment techniques, in the context of digital preservation and, more concretely, in the preservation of e-Science data and processes, in order to develop a Risk Assessment method that can be applied while managing the life-cycle of digital information.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: System Issues

General Terms
Management, Measurement.

Keywords
Risk Management, Risk Assessment, Digital Preservation, e-Science.

1. INTRODUCTION
Risk can be seen as the effect of uncertainty on objectives [2]; it is usually quantified as the combination of the probability of occurrence of an event and its consequences. Risk is everywhere and in everything we do, therefore, it is thoroughly necessary to rely on Risk Management (RM) to help us perceive and control risks. RM is constantly evolving and follows specific processes that can be applied to several contexts. Generic standards [1], [2], [3] can point us in the right direction when dealing with risk. However, one must keep in mind that, even though these standards can guide us in the right direction, they cannot give us an universal approach to RM, since every case is unique and has a different background.

Digital preservation (DP) is a blooming concern. Projects are being developed worldwide towards reaching the goal of maintaining digital objects (and the information they contain) accessible and understandable to users for long periods of time, and all the while making sure that both the integrity and the authenticity of these objects are upheld. To reach that, careful planning must be put in practice, clear objectives on which information to preserve and what level of protection it needs must be considered and the characteristics of the preservation environment must be established.

The achievement of DP objectives is a process, since there are numerous threats and vulnerabilities that can affect the ultimate objective of digitally preserve objects. Moreover, it also encloses several challenges to the preservation process itself, so, it needs a firm and trustworthy way to assess and treat the involved risks.

These risks increase when considering data and processes in the e-Science (or enhanced science) context. E-Science represents an alliance between science and IT; it is a collaborative and data-intensive approach, which comprises, besides the data itself, the technological infrastructure to support such huge amounts of information [9]. This is a growing area, and a growing reference on how to make scientific discoveries as well. It is collaborative science, and, consequently, deals with both large and complex raw data sets and information collections. As such, obtained data and employed processes must be digitally preserved for future reference, and this information's life-cycle must be thoroughly managed. Thus, the need for a comprehensive and methodological way to assess risks in this type of initiatives is a critical concern.

The worked presented in this paper was developed with the purpose of achieving a methodological way to assess risks in DP and, specifically, in the DP of e-Science data and processes. It went through understanding which risk assessment techniques are adequate in this context, and how they can be used and combined in order to reach a thorough method to apply known risk assessment techniques to this particular domain. The resulting risk assessment method can be, in the future, combined with DP techniques, meant to treat the assessed risks.

This paper is structured as follows. Section 2 outlines related approaches and standards in the areas of RM, DP and e-Science data and processes. Section 3 limits the problem addressed in this paper, while Section 4 presents the proposed approach to assess risks in the digital preservation of e-Science data and processes.

iPRES2011, Nov. 1–4, 2011, Singapore.

Finally, Section 5 lists the main conclusions of the presented research work.

2. RELATED WORK

The major areas of RM, DP and e-Science converge in the work presented in this paper. We discuss the main approaches and standards adopted in each area to provide an overview of their body of knowledge.

2.1 Risk Management

On a daily basis, we are presented with challenges, there is always a certain degree of uncertainty and even a previously established system, process, activity or operation can be exposed to new and emerging threats and vulnerabilities that could compromise our objectives. This is the very definition of risk (see Figure 1), the effect of uncertainty on previously set of objectives, combining the probability of an event's occurrence and the consequences it may cause.

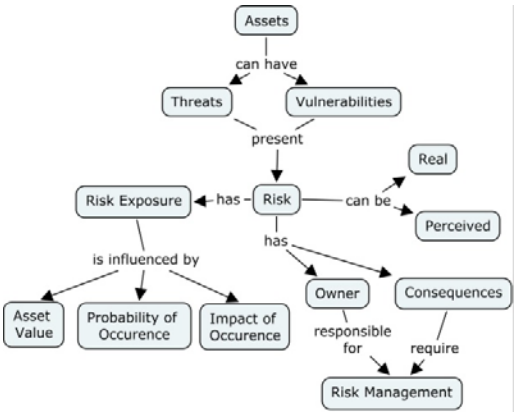


Figure 1 – Need for Risk Management

RM, which can be defined as a set of coordinated activities to direct and control an organization with regard to risk [1], and whose main goal is to define prevention and control mechanisms to address the risks attached to specific activities and valuable assets [4], should therefore be considered as an essential part of every organization and every project it may take on. RM should be iterative, not only applied while developing a project but also while operating and maintaining the resulting product [5], making sure changes that emerging risks are properly addressed.

Several standards exist in the scope of RM. Probably the most relevant of these standards is the ISO 31000:2009 [1], a set of principles and guidelines that can be used by "any public, private or community enterprise, association, group or individual" [1] when dealing with risk. It has two supporting standards as well: the ISO/IEC 31010:2009 [3], a standard guide describing systematic techniques for risk assessment; and the ISO Guide 73:2009 [2], a guide containing definitions for vocabulary terms related to RM.

Even though there are other prominent standards in this arena, like COSO ERM [10], AIRMIC, ALARM, IRM (AAIRM) [12], M_o_R [11], ISO/DIS 21500 [15], ISO 28000:2007 [16], Value-at-Risk [14], IT Governance Institute's Risk IT Framework [13], and OCTAVE [17], among others, the ISO 31000:2009 is the internationally recognized RM standard; thus, the work presented in this paper is mainly directed by the principles, concepts and guidelines provided in this standard family.

In order to guarantee a successful RM, a systematic RM process (see Figure 2) should be followed, in order to realize not only what the possible risks are, but also to analyze, evaluate and treat them, as well as to establish the context and criteria against which they should be judged. This process must be constantly monitored and reviewed in order to act on possible emerging risks; stakeholders must also be constantly involved in the process.

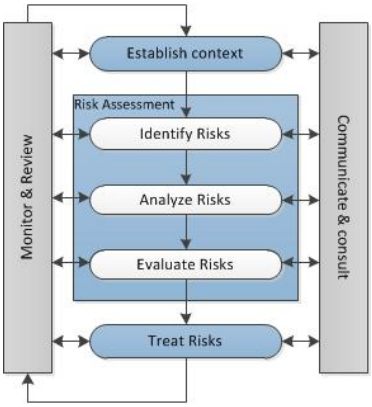


Figure 2 – Risk Management Process [1]

Perhaps the most important task of the whole RM process is risk assessment; and this is the focus of this paper. Risk assessment is not an easy task, it can be very subjective, has a strong dependency from the context where it is to be applied, and has to be a balance between science and judgment and take several psychological, social, cultural and political factors into account [6], which makes it a multidimensional problem. It can be done in either a quantitative, semi-quantitative or qualitative manner and should be as thorough as possible since, if the assessment fails, the subsequent risk treatment will also be inadequate, which may have catastrophic implications.

Assessing risks consists on identifying, analyzing and evaluating them. Risk identification involves ascertaining which events may occur that will jeopardize the normal behavior and/or development of a certain project or activity.

The goal of risk analysis is to understand the identified risks, through a multi-level analysis. There are three main views to risk analysis [3]: the consequence of the risk; the probability that the risk will occur; and the level of risk (combination of its consequences and probability).

The final stage of risk assessment is risk evaluation, where all the information gathered on the previous stages is used, along with the list of criteria produced when establishing the context, to prioritize risks and decide whether or not treatment is necessary.

Several methods and techniques can be used by Risk Assessment. The ISO/IEC 31010:2009 [3] standard surveys 31 techniques to perform Risk Assessment, and shows how they can be applied to each step of the Risk Assessment process as follows: (*i*) risk identification; (*ii*) risk analysis – consequence analysis; (*iii*) risk analysis – qualitative, semi-quantitative or quantitative probability estimation; (*iv*) risk analysis – assessing the effectiveness of any existing controls; (*v*) risk analysis – estimating the level of risk; and (*vi*) risk evaluation.

2.2 Digital Preservation

The main goal of DP is to provide long term preservation and accessibility of digital objects, while maintaining their authenticity and integrity [4].

Throughout time, important information, knowledge and data arise in a digital form, which must not be lost and should, therefore, be preserved for future use (see Figure 3). However, DP poses some serious problems, since not only the original content needs to be maintained, but one must be able to provide evidence that it is authentic, correct and has not been changed.

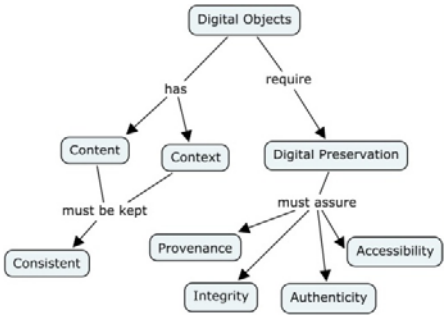


Figure 3 – Preservation Needs

DP aims at preserving digital objects for the long term, making sure the needs of future users are satisfied [7], allowing not only the ingestion and preservation of data, but also its dissemination, making it available to those whom it might concern. Since each type of digital object has its own specific set of requirements, this poses a great challenge, demanding an accurate planning of DP activities.

A common DP environment encompasses all the information entities, the control processes for those entities and the technological infrastructure to support the environment. However, the development of this environment is not a simple chore; not every repository is trustworthy enough to keep such sensitive items and preserve them for the long term, controlling the threats and vulnerabilities involved.

Such a repository must be reliable so as to keep the digital objects intact, accessible and authentic; it must also be flexible, scalable and heterogeneous, as to respond and adjust to emerging changes.

These concerns and requirements should all be taken in consideration while planning the DP process; there needs to be constant monitoring and auditing of this planning process, to make sure the DP plan is adequate to the established goals and requirements, and to make it possible to react to changes whenever they occur. Such monitoring and audit should also be a part of the DP process itself as to keep existing threats and vulnerabilities under control and to discover emerging ones as well, making sure we can timely and adequately react to every new change and challenge.

DP is very challenging to plan and undertake; it has many variables and perspectives to take in consideration. Hand to hand with the challenges come threats and vulnerabilities.

Even though everything is exposed to threats, and everything has vulnerabilities, when it comes to DP, this exposure may be especially dangerous, since we are dealing with information that can be a very sensitive, valuable and powerful asset. This is why these vulnerabilities (see Table 1) and threats (see Table 2) must be assessed from the very planning of the DP venture.

To help in this process, and even though there is not a standard way to approach DP, there are some standards and references that provide principles and guidelines for several steps of the process. The most prominent initiative addressing DP through RM is DRAMBORA [8], which is based on a generic RM process to propose a methodology for self-assessment, encouraging organizations to establish a comprehensive self-awareness of their objectives, activities and assets before identifying, assessing and managing the risks implicit within their organization.

Table 1 – Digital Preservation Vulnerabilities [4]

	Vulnerability	Description
Process	Software faults	bugs that can cause abnormal behavior or even software failure
	Software obsolescence	software becomes obsolete and unable to run or communicate with other components
Data	Media faults	irreversible bit errors (bit-rot) or irrecoverable loss of bulk data (disk crashes or loss of offline media)
	Media obsolescence	representation formats become obsolete and cannot be rendered
Infrastructure	Hardware faults	transient recoverable failures (power loss) or irrecoverable failures (burnt-out power supply unit)
	Hardware obsolescence	hardware becomes obsolete and unable to communicate with other components
	Communication errors	occur while transferring data, these errors might be detected but might also, in some cases such as check-sum errors, go by undetected
	Network services failures	such as DNS and persistent URL errors

Table 2 – Digital Preservation Threats [4]

	Threat	Description
Disasters	Natural disasters	such as earthquakes, floods and fires
	Human operator error	can include both recoverable and irrecoverable errors, such as data deletion; might also involve hardware or software components
Attacks	Internal attacks	malicious users, with privileged access to the organization or physical location of components, may cause: data orcomponent destruction or modification; denial of service; theft
	External attacks	similar to the internal attacks but done over public networks connections; may also encompass attacks such as viruses and worms
Management	Economic failures	budgets are not very stable when it comes to digital preservation, funding may become insufficient over time
	Organizational failures	such as political changes, incompetent management or other unpredictable reason; may lead to changes in what concerns digital preservation requirements, constraints, priorities, …
Legislation	Legislative changes	current processes for digital preservation or preserved data may not obey to the new or revised legislation
	Legal requirements	current processes for digital preservation, preservation environment, repository, and preserved data must obey to the current legislation; if not, legal punishments and fines may take place

2.3 e-Science data and processes

E-Science, which goes through several stages, (see Figure 4) takes science to a new paradigm, a collaborative one, which relies very much on data intensive computing and on community access to distributed data [9].

This new science paradigm comes with a whole new set of challenges, which derive mostly from the colossal amounts of data involved and the ability to share one's scientific information (whether raw data captured from sensors, instruments and/or simulations, or data analysis) and to view and use information shared by other scientists.

Many of the captured scientific data can be unrepeatable (it can be too costly to retake an experiment or even impossible due to external conditions and events); which would make losing that data a potential catastrophe, not only making it impossible to use that same data for further studies, but also any other data derived from it, since it would not be possible to attest to its provenance and authenticity.

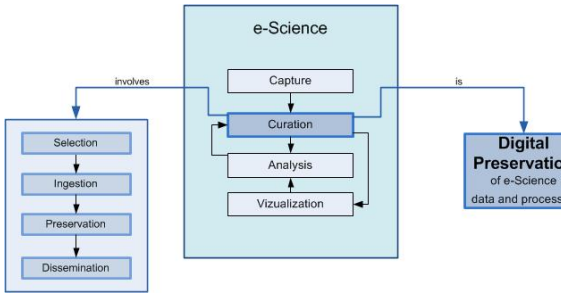


Figure 4 – e-Science activities

The sharing and collaboration aspect of e-Science poses several major issues; one of them is intellectual property. In such an environment, there are those who generate the original data, those who analyze it (possibly generating other resulting data as well), those who use it for research, etc., making it imperative to know where the data came from and who is responsible for it.

Since different analysis methods, workflows and processes can lead to different results and data and, if these methods and processes are not maintained and properly related to the corresponding data, that can lead to potentially mislead research and even misinformed decision making. Along with these workflows, processes and methods, logbooks regarding each experiment (if they are kept) must be duly related to the corresponding information as well.

When considering a digital repository containing e-Science information, one of the main issues is the quality of that information; one expects it to be correct, reliable and trustworthy enough to be useful in research and for further studies and analysis [9].

Though all general DP needs and requirements are maintained in this context, it poses even more demands and requires even more care, since the information might be the target of further exploitation and developments, and is not only meant to be read and consulted in the future.

3. PROBLEM CONTEXT

The information resulting from e-Science processes and workflows has a long life-cycle, which needs a very careful management, in order to assure the properties as well as the content of the information in question.

The DP arena developed several knowledge and best practices, but those concepts have been mainly applied to the cultural heritage sector. The e-Science domain imposes new requirements and raises several challenges on the way this problem should be addressed. In fact, while DP is the main driver of cultural heritage organizations, it must be addressed as an issue (among several other requirements) of the overall e-Science environment, where RM can be seen as a powerful approach to address the potential threats affecting the achievement of DP.

When digitally preserving e-Science information, most of the technological requirements are the same as general digital preservation ones. However, these scenarios come along with the necessity of standard formats and representation, to guarantee future understandability, and make sure the preserved information can be read and used by others in future studies, which also entails the preservation of processes along with the data objects. Also to make it possible for the preserved information to be used in future studies, there is the need to keep a more thorough context than a simple hardware and software one; it is necessary to keep experiments contexts (input parameters, etc.) for them to be able to be reproduced or validated.

While technological requirements of digital preservation are mostly maintained when dealing with complex e-Science scenarios, when it comes to the trustworthiness of the information, the requirements are more specific and require even more attention.

Before any data is ingested, there is the need to make a methodical selection, including a thorough validation of this data to assure no "bad" information, which might potentially taint studies and analysis, is preserved.

The need for authenticity assurance grows even larger when dealing with scientific information, it is absolutely imperative to be sure that a digital object corresponds to the information provided by the original owner, so as to make sure that no information contained in the repository is illegitimate and that digitally preserved data and processes actually correspond to those captured and/or used by scientists. For similar reasons, it is also strictly necessary to attest to the information's integrity for as long as it is preserved, guaranteeing no changes have been made to the informational content.

This need for integrity assurance is all the more pressing when dealing with this type of information, since ingested scientific data should never be subject to change. If the preserved information is used, and changes/additions are made, another version of that information must be ingested and appropriately related to the original one, in order for it to be able to be verified or even reused in the future. No scientific information, regardless of following developments, should be lost or written over, not even in case of discovered errors, bugs, etc., since it might be needed for future consultation or use.

It is necessary that the preserved information is absolutely correct, maintaining these properties, in order for data to be able to be used in further studies, analysis, and experiments or for processes and workflows to be reproduced, for example to confirm results and replicate experiments.

However, some of this information may not be supposed to be accessible for the general public, being restricted to certain entities or communities. Thus, it is necessary that some degree of confidentiality is maintained.

Long-term provenance is imperative to be kept, in order to guarantee not only the ability to identify who is responsible for the information but also intellectual property rights which are obviously important when it comes to scientific discoveries. These properties must be kept not only for captured data, but also for corrections (new versions) made to those data, and data analysis processes, workflows, and results, which may lead to

scientific breakthroughs and must, therefore, be associated with their rightful owners.

These analysis processes and workflows need also to be associated with the original data, as well as posterior results and, in case they are kept, logbooks, each with their own provenance assured, in order to guarantee intellectual property rights of each are maintained along the scientific information's life-cycle.

And this is a very long life-cycle: data and analysis results and processes are not only kept for consultation but can also be the subject of further analysis or studies and, even though the original information is never changed, new and associated information will keep rising.

For the digital preservation of e-Science data and processes to be successful, it is necessary to guarantee that these requirements and needs are met, which makes it imperative to manage possible risks in the most effective and possible way.

However, the use of RM methods in DP is still immature, and there is a lack of guidance to bring and apply the established RM concepts to the DP arena. In fact, despite DRAMBORA [8], a standard way to apply RM to DP does not exist; which would be an added value to the process of preservation, since it could provide specific methods to identify, analyze, evaluate and treat the risks presented in this process, which is becoming more vital with each passing day.

One of the most important phases of RM is Risk Assessment, which consists on identifying, analyzing and evaluating potential risks. Risk Assessment is completely vital to RM in general and DP in particular, since, if the assessment of risks fails, the subsequent treatment will most likely be inadequate, causing the failure of the whole RM process. As such, and, since it is a very complex and extensive area on its own, risk assessment is the main focus of this paper, leaving the treatment of risks as future work.

Since science has always and will always play such a big and important role, a thorough Risk Assessment of e-Science digital repositories is essential. This was one of the main drivers of this work.

Thus, we propose a method to guide Risk Assessment in DP of e-Science data and processes. Its main focus lies on the management of the information's long life-cycle, and it is meant to provide a way to, given a specific scenario in this particular domain, be able to detect and quantify potential threats. This approach can be seen as a complement to generic RM processes or the DRAMBORA approach to DP. It is not an alternative, but a guide for the Risk Assessment activities in DP.

4. PROPOSED APPROACH

The proposed Risk Assessment method was developed through the comprehensive study of known risk assessment techniques (see Figure 5); this study was mostly based on [3] and is meant to complement DRAMBORA [8].

A previous separation of Risk Assessment techniques was made, dividing them into identification techniques, analysis techniques and evaluation techniques, according to which of these Risk Assessment activities they could be applied to. While all the identification techniques were studied with regards to their applicability to the DP context, both the analysis and evaluation techniques were further separated, in order to rule out those that, from the start, were not adequate to the creation of a complete Risk Assessment method. As such, the analysis and evaluation techniques were separated into representative and rating techniques, and the first ones were excluded (when applying a method in a systematic way, these techniques are too subjective, allowing for different interpretations and, consequently, possibly different results when applying this method to the same scenario, thus compromising the correctness of the method itself).

Afterwards, the rating techniques (which can be qualitative, semi-quantitative, and quantitative), along with all of the risk identification techniques, were subjected to a primary and general analysis in order to discard those techniques that, from the start, were not adequate to the scenario at hand. An example of such a technique is the Environmental Risk Assessment, whose scope (people, animals and plants) is completely divergent from the one of this work.

From that point, all the remaining techniques were studied and analyzed in detail, in order to establish their capability of correctly identifying risks (whether known or new), analyzing and evaluating them in each of the different DP of e-science data and processes activities. This was accomplished by verifying the compliance of these techniques with a list of objectives, needs and requirements imposed by the context at hand.

After the individual analysis of each technique was done, a more global study took place. Techniques were compared in order to ascertain the most suitable ones, to be applied in each of the DP of e-science data and processes stages and activities, among the existing possibilities; dependencies between techniques were studied to understand which of these it made sense to combine in each activity of the risk assessment. Even though other techniques may be used, and each case is always a different case, the techniques found in the next subsections are the ones that we recommend to be used in this type of scenario.

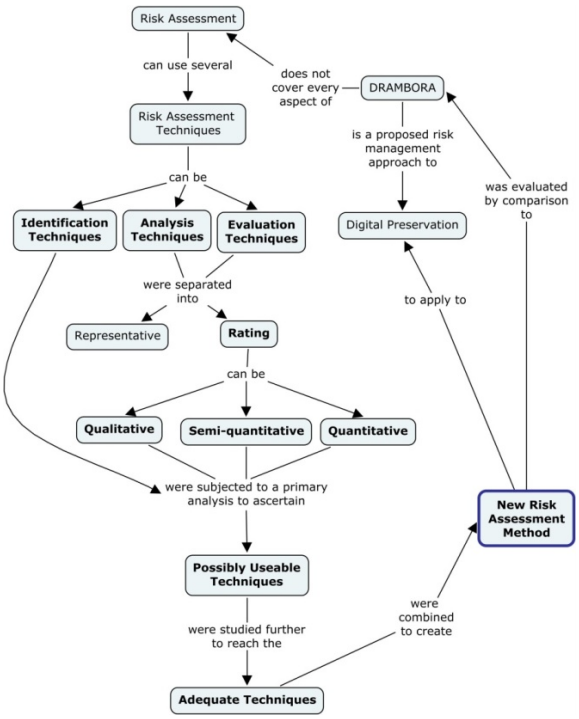


Figure 5 – Developed work

Table 3 – Risk identification techniques

	Context				Types of Risk					Recommended for risk identification
Risk Identification Technique	Problem context & scope	Feasible	Systematic	Comprehensive	Known	New	Human	System	Process	
Check-lists	√	√	√	×	√	×	√	√	√	Yes
PHA	×	–	–	–	–	–	–	–	–	No
Brainstorming	√	√	√	√	√	√	√	√	√	Yes
Interviews	√	√	√	√	√	√	√	√	√	Yes
Delphi Technique	√	×	√	√	√	√	√	√	√	No
SWIFT	√	√	√	√	√	√	√	√	√	Yes
Environmental Risk Assessment	×	–	–	–	–	–	–	–	–	No
Scenario Analysis	√	√	×	×	√	√	√	√	√	No
BIA	×	–	–	–	–	–	–	–	–	No
FTA	√	√	√	×	√	√	√	√	√	No
ETA	√	√	√	×	√	√	√	√	√	No
Cause-Consequence Analysis	√	√	√	√	√	√	√	√	√	Yes
Cause & Effect Analysis	√	√	×	×	√	√	√	√	√	No
CBA	×	–	–	–	–	–	–	–	–	No
MCDA	√	√	×	×	√	√	×	×	×	No
HAZOP	√	√	√	√	√	√	√	√	√	Yes
HACCP	√	√	√	×	√	√	×	×	√	Yes
FMEA	√	√	√	√	√	√	√	√	√	Yes
RCM	√	√	√	×	√	√	×	√	×	Yes
HRA	√	√	√	×	√	√	√	×	×	Yes
SA/SCA	√	√	√	√	√	√	×	√	×	No
LOPA	×	–	–	–	–	–	–	–	–	No
Markov Analysis	×	–	–	–	–	–	–	–	–	No
FN Curves	√	√	×	×	√	√	√	√	√	No
Risk Indices	√	×	×	√	√	√	√	√	√	No
Consequence / Probability Matrix	√	√	√	×	√	×	√	√	√	No

4.1 Risk Identification Techniques

A summary of the carried out analysis, regarding the risk identification techniques and their applicability to the problem context, can be found in

Table 3. The columns of this table have the following purposes:

- 1st column: shows whether or not the technique is applicable to the context at hand as well as to the project's scope;
- 2nd column: shows whether or not it is feasible/realistic the use of that technique in the context at hand (having in mind the possible constraints regarding resources, time, etc.);
- 3rd column: indicates if it can be applied in a systematic manner;
- 4th column: specifies whether the technique is comprehensive when it comes to the potential risks;
- 5th, 6th, 7th, 8th and 9th columns: regard the types of risk which can be identified through the use of that technique (known or new; of human, process, or system nature);
- 10th column: states whether or not it was recommended to be used in the scenario at hand for risk identification purposes.

After the study of all the risk identification techniques, these, in this order, are the ones that we propose to be applied to the DP of e-Science data and processes:

- **Check-lists**, as a preliminary technique, to provide a starting point to the identification of risks, and guarantee no known/common risks to digital preservation are overlooked;
- **Brainstorming**, using a formal process, to have a group of knowledgeable stakeholders gather a list of both known and new risks regarding the scenario at hand in a systematic manner;
- **Interviews**, to target specific stakeholders with the aim to identify "concern-related" risks, and provide further details on risks potentially related to those identified by check-lists and brainstorming;
- **Structured "what-if" technique (SWIFT)**, to be used when change is eminent, particularly taking into consideration the selection, preservation and dissemination stages of the curation process, where change can be more influential, to identify potential risks arising from that change;
- **Failure Mode and Effect Analysis (FMEA)**, to identify design objective deviations and associated risks, potential causes, and consequences, regarding both the curation process and the digital repository itself, while making sure digital preservation's objectives, needs, and requirements have not been neglected;
- **Reliability Centered Maintenance (RCM)**, used along with FMEA, resorting to a specific approach to the latter, in order to identify preventive measures and policies that should be put in place to protect the digital repository, especially regarding the ingestion, preservation, and dissemination phases of the curation process, which are the ones which rely on the repository;
- **Human Risk Assessment (HRA)**, to assess possible human impact on every stage of the curation process;

Table 4 – Risk analysis techniques analysis summary

	Context				Considers			Properties		Recommended for risk analysis
Risk Analysis Technique	Problem context & scope	Feasible	Systematic	Comprehensive	Probability	Consequence	Level of risk	Objective	Possibly Quantitative	
SWIFT	√	√	√	×	√	√	√	√	√	No
RCA	√	√	×	×	√	√	√	√	×	No
Environmental Risk Assessment	×	–	–	–	–	–	–	–	–	No
Scenario Analysis	√	√	×	×	√	√	√	×	√	No
BIA	×	–	–	–	–	–	–	–	–	No
FTA	√	√	√	×	√	×	√	√	√	No
ETA	√	√	√	×	√	√	√	√	√	No
Cause-Consequence Analysis	√	√	√	√	√	√	√	√	√	Yes
Cause & Effect Analysis	√	√	×	×	×	√	×	√	×	No
Decision Tree	√	√	√	√	√	√	√	×	√	Yes
CBA	×	–	–	–	–	–	–	–	–	No
MCDA	√	√	×	×	√	√	√	×	√	No
HAZOP	√	√	√	√	√	√	√	×	×	No
HACCP	√	√	×	×	×	√	×	√	√	No
FMECA	√	√	√	√	√	√	√	√	√	Yes
RCM	√	√	√	×	√	√	√	√	√	Yes
HRA	√	√	√	×	√	√	√	√	√	Yes
LOPA	×	–	–	–	–	–	–	–	–	No
Bow-tie Analysis	√	√	×	×	√	√	√	×	√	No
Markov Analysis	×	–	–	–	–	–	–	–	–	No
Bayesian Analysis	×	–	–	–	–	–	–	–	–	No
FN Curves	√	√	×	×	√	√	√	×	×	No
Risk Indices	√	×	×	√	√	√	√	√	√	No
Consequence / Probability Matrix	√	√	√	×	√	√	√	√	√	used in FMECA

- **Cause-consequence analysis**, to make sure possible underlying and/or consequent risks relating to the previously identified risks are not neglected. This technique can also be used to understand which risks are related among each other.

4.2 Risk Analysis Techniques

When it comes to risk analysis, a summary of the undertaken study regarding their applicability can be found in

Table 4. The columns of this table have the following purposes:

- First 4 columns: the same as those in Table 3;
- 5th, 6th, 7th and 8th columns: refer to whether or not probabilities, consequences, and/or the level of risk are considered by each risk analysis technique;
- 8th column: indicates if the technique can be objective, not giving room for different interpretations in the same situation;
- 9th column: states if the technique in question can be used quantitatively;
- 10th column: states whether or not it was recommended to be used in the scenario at hand for risk analysis purposes.

After the study of the available techniques, the ones that we propose to be used in the context of DP of e-Science data and processes, and the order in which they should be applied are:

- **Decision tree**, to be used considering the selection stage of the DP process, which is where most decisions are made, in order to estimate, for each path coming from a certain decision/event, the value/cost of its outcome, to provide means to later choose the best from the available set of options;
- **Failure Mode Effect and Consequence Analysis (FMECA)**, resorting to the use of a **consequence/probability matrix**, to calculate each risk's criticality, in order to both provide the means to later prioritize risks and serve as input to cause-consequence analysis;
- **Reliability Centered Maintenance (RCM)**, used along with FMECA, resorting to a specific approach to the latter, to estimate the frequency of each failure that may occur especially in the ingestion, preservation, and dissemination phases of the DP process, in case maintenance is not performed;
- **Human Risk Assessment (HRA)**, to calculate probabilities and possible consequences of human error in the DP process and provide input to cause-consequence analysis;
- **Cause-consequence analysis**, to analyze the possible causal and consequent risks of each of the identified risks, and calculate their probabilities and possible consequences.

4.3 Risk Evaluation Techniques

Regarding the study of risk evaluation techniques (a summary of this analysis can be found inTable 5, where the columns have the same meaning as the corresponding ones in

Table 4), the ones that we propose as most suitable to be used in the DP of e-Science data and processes are (in this order):

- **Decision tree**, to be used considering the selection stage of the DP process, which is where most decisions are made, and choose the best from the available set of options, taking into account the previously made analysis;

Table 5 – Risk evaluation techniques analysis summary

		Context				Properties		Recommended for risk evaluation
	Risk Evaluation Technique	Problem context & scope	Feasible	Systematic	Comprehensive	Objective	Possibly Quantitative	
6	SWIFT	√	√	√	×	√	√	No
7	RCA	√	√	×	×	√	×	No
8	Environmental Risk Assessment	×	–	–	–	–	–	No
9	Scenario Analysis	√	√	×	×	×	√	No
10	BIA	×	–	–	–	–	–	No
11	FTA	√	√	√	×	√	√	No
13	Cause-Consequence Analysis	√	√	√	√	√	√	Yes
15	Decision Tree	√	√	√	√	×	×	Yes
16	CBA	×	–	–	–	–	–	No
17	MCDA	√	√	×	×	×	√	No
18	HAZOP	√	√	√	√	×	×	No
19	HACCP	√	√	×	×	√	√	Yes
20	FMECA	√	√	√	√	√	√	Yes
21	RCM	√	√	√	×	√	√	Yes
22	HRA	√	√	√	×	√	√	Yes
25	Bow-tie Analysis	√	√	×	×	×	√	No
27	Monte Carlo Simulation	×	–	–	–	–	–	No
28	Bayesian Analysis	×	–	–	–	–	–	No
29	FN Curves	√	√	×	×	×	×	No
30	Risk Indices	√	×	×	√	√	√	No
31	Consequence / Probability Matrix	√	√	√	×	√	√	used in FMECA

- **Human Risk Assessment (HRA)**, to be used according to the previously made analysis, by realizing which errors or task failures have higher contribution to risk, so as to establish risk priorities and decide whether or not a risk should be treated;
- **Failure Mode Effect and Consequence Analysis (FMECA)**, resorting to the use of a **consequence/probability matrix**, to prioritize the previously analyzed risks, and decide whether or not they should be treated based on this prioritization;
- **Reliability Centered Maintenance (RCM)**, used along with FMECA, resorting to a specific approach to the latter, to prioritize risks according to the previously estimated frequency of each in case maintenance is not performed;
- **Cause-consequence analysis**, by using the previously analyzed fault trees, present in this analysis, in order to prioritize risks and decide on their treatment based on their estimated probabilities and consequences;

4.4 The Proposed Method

Finally, the chosen techniques and combinations were all put together to create a Risk Assessment method for the DP of e-Science data and processes. This method can be found in Figure 6.

This is a cyclic method, intended to be used as a guide, which allows for the overall assessment of risk in this domain, focused on providing a tool to help in the management of this information's life-cycle, supplying the means to identify, analyze and evaluate risks throughout this life-cycle.

The proposed method follows the risk assessment phase of the RM process (see Figure 2).

It starts by identifying risks, where the proposed techniques can be used either separately or together (if used together they should follow the proposed order) and result in a preliminary set of documents encompassing a list of identified risks and some attributes of these risks, as well as documents resulting from the used techniques, such as diagrams, tables and figures.

When risk identification is done, risk analysis takes place and, again, the proposed techniques can be used either separately or together and, if used together, they should follow the proposed order; risk analysis results in an intermediate set of documents, including a more complete risk list, with some more attributes, and documents resulting from the used techniques, including diagrams, tables, figures and necessary calculations.

These documents will then be the input to the risk evaluation stage, which, as the previous two stages, can be done through the use of the proposed techniques (either separately or together), and uses the given inputs to decide whether or not the identified and analyzed risks should be treated; this results in a final risk list.

Thus, as an output, besides the intermediate documents containing the figures and results from each of the three risk assessment activities, this method provides a document containing a list of risks, encompassing, for each risk, a set of attributes to describe it.

These attributes encompass: the risk's nature (whether it's a system risk, process risk, human-related, etc.); the phase(s) of the DP process where the risk may arise; the techniques used to assess the risk; the risk owner (the one responsible for it, from the moment it is identified); affected stakeholders; related risks; probability of occurrence; risk's consequence (only regarding the potential loss of digital objects); the resulting level of risk (combination between the probability and consequence); the date its assessment was completed; the risk's priority.

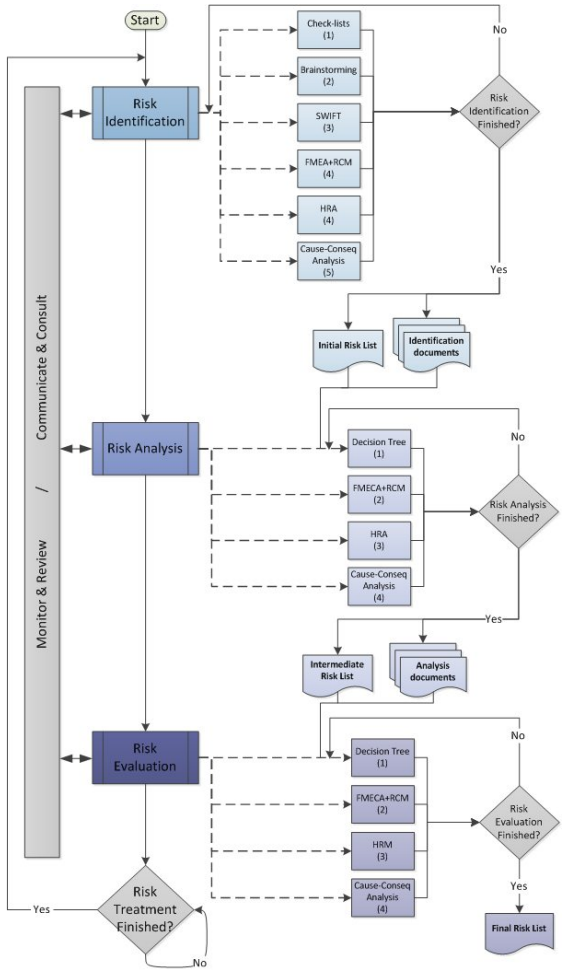


Figure 6 – Proposed Risk Assessment Method

This method provides a comprehensive way of assessing risks in scenarios of DP of e-Science information, from their identification, to their analysis and evaluation. It provides guidance when it comes to the more suitable risk assessment techniques to be used in this context, along with how they may be combined to be as complete and thorough as possible, indicating the best means to aid in identifying a broader range of risks (regarding all the elements involved in DP), analyzing and evaluating them.

4.5 Results and Evaluation

To evaluate the proposed method, a concrete e-Science scenario was used. This scenario concerns LIP[1], a scientific and technical laboratory of particle physics.

A commonly used software to simulate experiments in the high energy physics and astroparticles arena is CORSIKA[2], which is a modular program and requires that each different simulation follows a specific process (see Figure 7).

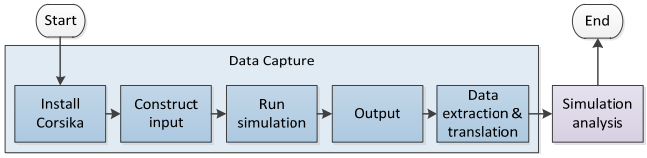


Figure 7 – CORSIKA simulation and analysis process

Several factors may influence this process's outcome, for instance:

- The decisions made can influence the entire process as well as the outcome;
- Changes in the CORSIKA software version may affect the simulation's output;
- Different parameters in CORSIKA installation may affect the simulation's output;
- Different options in the input file may affect the simulation's output;
- Different translations (possible due to the ambiguous manual) may originate different data;
- Different analysis may originate different data.

Some of these simulations can be too costly to reproduce (some run for a long time and have large outputs) and the generated program and outputs must be kept for as long as possible, in order to be able to use the data and verify conditions and results. Hence, the need for digital preservation arises and, along with it, a whole new set of risks.

Through the use of the proposed method, it is possible to assess risks in this particular scenario, in a comprehensive way, by identifying risks which are not as commonly found in known digital preservation risks (as those identified in DRAMBORA reports [8]), analyzing and evaluating them.

This specific case did not call for the use of all of the proposed risk assessment techniques, since it was a fairly simple scenario, to be considered prior to any preservation effort, strictly on a theoretical basis, for the time being. Thus, a simple technique could be used and have a thorough result all the same.

As such, the following examples of possible risks were identified through brainstorming and analyzed and evaluated through FMECA (these risks' probabilities, consequences and levels of risk are represented in Figure 8 by means of a consequence/probability matrix):

R1 – Loss of data translation information, a system/process risk which can arise during the preservation or dissemination stages of the DP process and affects those wanting to analyze data. Since this risk was categorized as Level III, it should be treated as soon as possible.

R2 – Loss of relationship information between preserved analysis processes/workflows and the original data, a system/process risk which can arise during the preservation or dissemination stages of the DP process and affects future results confirmation. Since this risk was categorized as Level II, it should be monitored to see if the risk escalates, in which case treatment might be needed.

R3 – Loss of CORSIKA input parameters for a given simulation, a system/process risk which can arise during the preservation or dissemination stages of the DP process and affects future simulation recreation. Since this risk was categorized as Level II, it should be monitored.

R4 – Selection of incorrect information due to erroneous data validation, a human/process risk which can arise during the

[1] http://www.lip.pt

[2] http://www-ik.fzk.de/corsika/

selection stage and influence all the future stages of the DP process and affects every analysis, study or consultation made based on that information. Since this risk was categorized as Level IV, it should be treated immediately.

R5 – Loss of CORSIKA software version information, regarding a given simulation, a system/process risk which can arise during the preservation or dissemination stages of the DP process and affects future simulation recreation. Since this risk was categorized as Level II, it should be monitored.

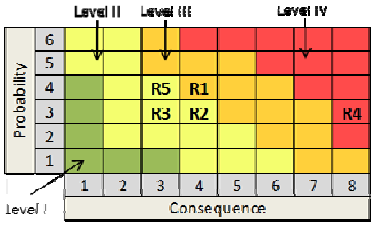


Figure 8 – Levels of risk

Even though these risks can be associated with known needs and requirements of DP (see section 2.2), their specificity prevents them from being listed in commonly used general DP risk checklists.

DRAMBORA [8] provides guidelines regarding digital repositories in general, proposing a methodology for self-assessment, encouraging organizations to establish a comprehensive self-awareness of their objectives, activities and assets before identifying, analyzing and managing the risks implicit within their organization. It has a risk management approach to digital preservation to assess and audit digital repositories.

However, since this is a general approach, to be applied to several digital repositories, it lacks the ability to extend to specific scenarios and, thus, to identify and further assess some unknown/uncommon risks that may rise in these cases.

This is especially evident when it comes to risk identification; since this is the starting point or risk assessment, it should be as thorough and comprehensive as possible. The proposed method recommends several techniques to be used in risk identification, providing a way to identify a wide range of risks instead of only those present in a general check-list.

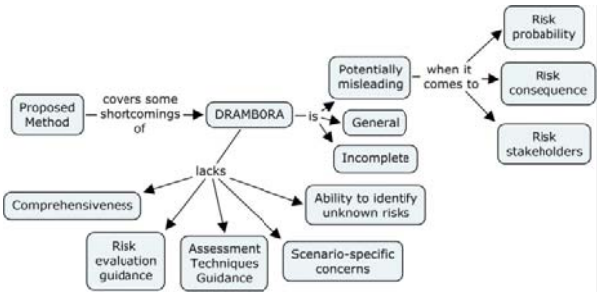


Figure 9 – Proposed method as complement of DRAMBORA

One of the biggest shortcomings in DRAMBORA is exactly this lack of guidance regarding risk assessment techniques. The suggested check-list for risk identification and the level of risk calculation for risk analysis may, in fact, be sufficient to assess some general, common repositories; however, when it comes to more specific cases, and especially when it comes to risk identification, they can be incomplete and not as comprehensive as they should be.

Moreover, the risk list example provided by DRAMBORA, which presents the results of both risk identification and analysis, may be misleading in some cases, given that it lists the same stakeholders, probabilities, and consequences to every single risk.

Another very important shortcoming of DRAMBORA, in what concerns risk assessment, is the lack of guidance when it comes to risk evaluation, not giving any basis on risk prioritization and decisions concerning whether or not to proceed to risk treatment. In fact, this methodology considers only two risk assessment tasks as part of the whole risk management process: “identify risks” and “assess risks”, being that the latter corresponds simply to risk analysis.

All this comes to show the need of a more comprehensive and scenario-specific risk assessment method, as the one proposed in Section 0. This method allows for the identification, analysis and evaluation of both known, general, and new, more particular, risks, that can only be identified through techniques which can be adapted to a given scenario; in this case, the DP of e-Science data and processes.

Thus, the proposed method can be used as a complement to DRAMBORA (see Figure 9), covering its shortcomings, and serving as a guide for each of the three risk assessment’s phases.

5. CONCLUSIONS

RM is an ever evolving area, with application in numerous areas of our lives, businesses and organizations; there is always room for innovation, for creating new and better ways to address risks. Since e-Science’s collected and processed data and used proceedings should be able to be used for future consultation and reference, they must be digitally preserved. This imposes an immense set of risks, regarding both the handling of the data itself and its preservation.

This paper proposes a Risk Assessment method to guide in such efforts, by providing a systematic and comprehensive approach to identifying, analyzing and evaluating potential risks.

The particle physics evaluation scenario, to which the proposed method was subjected, encompassed all 3 stages of Risk Assessment regarding DP. Risk monitoring, communication and treatment tasks fall outside of this work’s scope; however, since the analyzed risks can be mapped into the risk taxonomy presented in section 2.2, this work provides a decision support basis when it comes to evaluating risk treatment controls as well.

The main goal of this approach is to identify wide-ranging risks (regarding the whole DP process, infrastructure, etc.), instead of focusing the assessment strictly on a component level, in order to identify as many risks as possible, to then analyze and evaluate.

6. 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.

7. REFERENCES

[1] ISO/FDIS 31000 (2009) *Risk Management - Principles and guidelines.*

[2] ISO Guide 73. (2009) *Risk management - Vocabulary*.

[3] ISO/IEC 31010 (2009) *Risk management - Risk assessment techniques*.

[4] Barateiro, J., Antunes, G., Freitas, F., Borbinha, J. (2010) Designing digital preservation solutions: a Risk Management based approach. *The International Journal of Digital Curation, Issue 1, Vol. 5*, 2010.

[5] Boehm, B. W. (1991) Software Risk Management: Principles and Practices, *IEEE Software, Number 1, Vol. 8*, 1991.

[6] Slovic, P. (2001) The risk game. *Journal of Hazardous Materials, Issue 86*, 2001.

[7] Doyle, J., Paquet, E., Viktor, H. L. (2007) Long term digital preservation - An end user's perspective. *2nd International Conference on Digital Information Management*, 2007.

[8] McHugh, A., Ruusalepp, R. Ross, S. & Hofman, H. (2007). The Digital Repository Audit Method Based on Risk Assessment. *DCC and DPE, Edinburgh*. 2007.

[9] Hey, T., Tansley, S., Toll, K. (2009) The Fourth Paradigm – Data-Intensive Scientific Discovery, *MS Research*, 2009.

[10] Committee of Sponsoring Organizations of the Treadway Commission (COSO) (2004) Enterprise Risk Management — Integrated Framework, *Jersey City, NJ: AICPA*, 2004.

[11] Office of Government Commerce (OGC) (2007) Management of Risk: Guidance for Practitioners (M_o_R), *United Kingdom*, 2007.

[12] Association of Insurance and Risk Managers (AIRMIC), ALARM (National Forum for Risk Management in the Public Sector), Institute of Risk Management (IRM) (2002) A Risk Management Standard, *London*, 2002.

[13] IT Governance Institute (2009) The Risk IT Framework.

[14] Holton, G. (2003) Value-at-Risk: Theory and Practice, *Academic Press*, 2003.

[15] ISO/DIS 21500 (2011) *Guidance on Project Management*.

[16] ISO 28000 (2007) *Specification for security management systems for the supply chain*.

[17] Caralli, R. A., Stevens, J. F., Young, L. R., Wilson, W. R. (2007) OCTAVE Allegro: Improving the Information Security Risk Assessment Process, *Software Engineering Institute at Carnegie Mellon University*, 2007.

Using Grid Federations for Digital Preservation

Gonçalo Antunes
IST/INESC-ID Information Systems Group
Lisbon, Portugal
goncalo.antunes@ist.utl.pt

Helder Pina
IST/INESC-ID Information Systems Group
Lisbon, Portugal
helder.pina@ist.utl.pt

ABSTRACT
Digital preservation aims at guaranteeing that data or digital objects remain authentic and accessible to users over a long period of time, maintaining their value. Several communities, like biology, medicine, engineering or physics, manage large amounts of scientific information, including large datasets of structured data that matters to preserve, so that it can be used in future research. To achieve long-term digital preservation, it is required to store digital objects reliably, preventing data loss. The data redundancy strategy is required to be able to successfully preserve data. Many of the characteristics required to implement, manage and evolve a preservation environment are already present in existing data grid systems, such as replication and the possibility to federate with other grids in order to share resources. We propose the customization of a data grid platform in order to be able to take advantage of its replication and federation features. In that way, scenarios where federated grids not thought for preservation purposes can be extended to preservation and their spare resources used with that mission.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Systems Issues; H.3.4 [**Systems and Software**]: Distributed Systems

General Terms
Algorithms, Design, Reliability, Verification.

Keywords
Data Grids, Digital Preservation, Redundancy, Federations, Replication.

1. INTRODUCTION
The Institute of Electrical and Electronics Engineers (IEEE) defines interoperability as 'the ability of two or more systems or components to exchange information and to use the information that has been exchanged' [1]. Digital preservation aims at ensuring interoperability in the time dimension (interoperate with the future), that is, guarantee that data or digital objects remain authentic and accessible to users over a long period of time, maintaining their value.

iPRES2011, Nov. 1–4, 2011, Singapore.

Several communities, like biology, medicine, engineering or physics, manage large amounts of scientific information. It usually includes large datasets of structured data (e.g., data captured by sensors), physical or mathematical simulations and several highly specialized documents reporting the work and conclusions of researchers.

The above mentioned information can be represented in a wide range of file formats and include a high level of relations that are not expressed in the data model of the file format. Moreover, the collaborative environment of the scientific community, and associated services and infrastructures, usually known as e-Science (or enhanced Science) [2], involves the requirement of interoperability and the respective data sharing. In a broad sense, 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. It also means, on the same time, a need and an opportunity for a better integration between science and engineering processes. Thus, long-term preservation can be thought as a required property for future science and engineering, to assure communication over time, so that information that is understood today is transmitted to an unknown system in the future.

In order to successfully transmit information to future generations, several strategies are possible such as format or storage media migration, emulation of hardware and software environments to be able to render the information, hardware refreshing, inertia, preservation metadata, and auditing [3].

To achieve long-term digital preservation, it is required to store digital objects reliably, preventing data loss. One potentially relevant strategy to achieve this goal is combining redundant storage and heterogeneous components. In using the redundancy strategy, digital preservation systems can take advantage of a basic attribute of digital information: it can be copied without any loss of information. This means that several copies of the data can be stored across many components. Through the use of the diversity strategy, which promotes the diversification of the properties of the components, the number of simultaneous failures in the system can be limited and the system is more likely to survive to a large correlated failure, such as in the case of a worm outbreak.

Achieving the goal of digital preservation may require a large investment in infrastructure for storing data, and on its management and maintenance. Such costs may be prohibitive for small organizations, or organizations that do not have steady revenue, like university libraries, research laboratories, or non-profit organizations.

An already common low-cost technology to handle e-Science collaboration and data management is the use of data grids [4].

Table 1 – Digital preservation threats and vulnerabilities taxonomy [3]

Vulnerabilities	Process	Software faults Software obsolescence
	Data	Media faults Media obsolescence
	Infrastructure	Hardware faults Hardware obsolescence Communication faults Network service failures
Threats	Disasters	Natural disasters Human operational errors
	Attacks	Internal attack External attacks
	Management	Economic failures Organization failures
	Legislation	Legislation changes Legal requirements

These are highly relevant solutions for digital preservation, as they already store massive amounts of the data that must be preserved, such as in e-Science domains, and they provide a set of functionalities required by digital preservation systems (e.g., redundancy, diversity). Furthermore, grids can be organized in different ways [5]. In particular, grids can be federated with each other. The federation model allows grids belonging to different institutions, and thus with independent administration and in different locations, to interoperate with each other so that data can be shared.

The iRODS data grid [6] is an adaptive middleware system that facilitates the management of data and policies according to the needs of the users. For that, it uses a rule engine to enforce and execute adaptive rules. Additionally, it supports the federation of different iRODS deployments. However, iRODS is not addressing specific digital preservation requirements, requiring customization to do so.

We propose the customization of the iRODS data grid platform in order to be able to take advantage of two types of scenarios: (*i*) grids *exclusive* for preservation, which comprises machines dedicated to running the data grid exclusively for digital preservation, which are likely to be under administration of the data owner; and (*ii*) grids *extended* for preservation, in which existing grid clusters, initially created for data processing, can be federated through the installation of an iRODS instance and extended for preservation. Their spare disc space, CPU, and bandwidth can be used to store data according to the preservation requirements. To be able to do so, we propose a set of micro-services and rules that use the replication features and federation configurations of iRODS to maintain data replicated geographically, so that it can be preserved from threats.

This paper is organized as follows. Section 2 describes related work, such as digital preservation threats and vulnerabilities, Data Grids, the iRODS data grid system, and the usage of data grids for preservation purposes in previous projects or publications. In section 3 we describe the problem of configuring the iRODS data grid in order to be able to take advantage of its replication and federation features. Then, in section 4 we discuss our proposal, and we finally conclude in section 5.

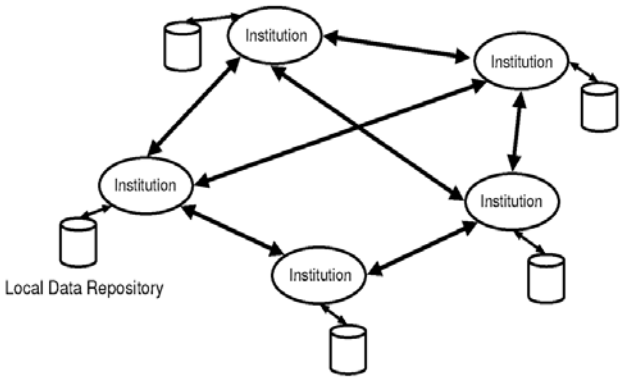


Figure 1. Grid federation model [5]

2. RELATED WORK

In this section, we describe related work, such as digital preservation threats and vulnerabilities, data grids, the iRODS data grid system, and the usage of data grids for preservation purposes.

2.1 Digital Preservation Threats and Vulnerabilities

In [3], the authors define a taxonomy of digital preservation threats and vulnerabilities in which a preservation environment is considered the aggregation of different components, namely the information entities, including preserved objects and metadata, processes controlling the information entities, and the technological infrastructure that supports the preservation environment.

Based on that assumption, each of these components may present several vulnerabilities: (*i*) process vulnerabilities, affecting the execution of processes (manual or supported by computational services) that control information entities; (*ii*) data vulnerabilities, affecting the information entities; and (*iii*) infrastructure vulnerabilities, enclosing the technical problems in the infrastructure's components.

Processes supported by software services can be affected by software faults and software obsolescence. Data vulnerabilities include media faults and media obsolescence. Infrastructure components can suffer hardware faults, hardware obsolescence, communication faults, and network services failures.

As for threats, those can be classified into disasters, attacks, management and legislation. Management failures are the consequences of wrong decisions that produce several threats to the preservation environment, such as economic failures and organization failures. Disasters correspond to non-deliberate actions that might affect the system, such as human operational errors, or uncontrollable events, such as natural disasters. Attacks correspond to deliberate actions affecting the system, such as internal or external attacks. Finally, legislation threats occur when digital preservation processes or preserved data violate existing legal requirements, or new or updated legislation (legislation changes).

2.2 Data Grids

Since it was defined in the 90's, many applications of this technology were made, and grids are used in scientific research projects, in enterprises, and other environments that require high

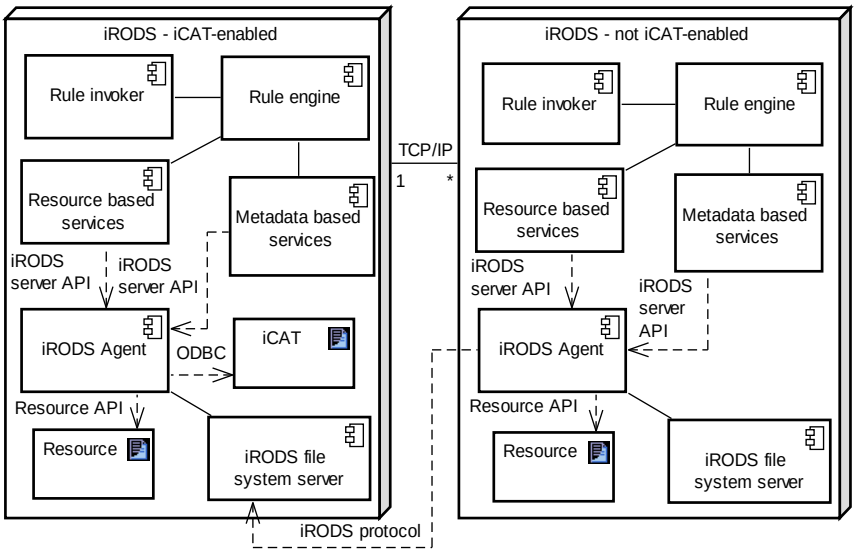


Figure 2 - iRODS deployment diagram [7]

processing power, while using low-cost hardware. Foster proposed a definition that consists in three properties that a system must comply with in order to be considered a grid [8]: (*i*) Resources are subjected to decentralized control; (*ii*) Standard, open, and general purpose protocols and interfaces are used; (*iii*) Nontrivial qualities of service are delivered.

In recent years, research has been done in defining a new type of grids, which deal with the management, sharing and processing of data. These were called data grids. Data grids o–er distributed services and infrastructure that allow the support of applications that deal with massive data blocks stored in heterogeneous distributed resources [9]. In data grids, data is organized as collections or datasets, and is replicated using a replica management system that creates, manages and modifies replicas. Information about replicas is organized in a replica catalog. The main characteristics are the following:

- Massive Datasets: Data grids allow the management and access to enormous quantities of data, in the order of terabytes or even petabytes [10].
- Logical Namespace: Is provided through the use of virtual names for resources, files and users. In the case of resources and files, one logical name maps to one or more physical names.
- Replication: Increases scalability and reliability through greater availability and redundancy. Data grids, as big distributed systems, must implement data replication mechanisms, in order to guarantee system scalability.
- Authorization and Authentication: Due to the high importance and frailty of some of the shared data, authentication and authorization mechanisms must be taken into account in order to comply with the authenticity and integrity requirements.

Grids can be organized in federated zones. Each zone has full control of its administrative domain and can operate independently of other zones. A federation of zones allows the sharing of data and resources between zones in the federation. The main benefits of this configuration are Location Transparency, as users can access resources at any node in a transparent way; Availability, as the replication in di–erent storage media, in di–erent locations allows the data to be available throughout the grid; Administration, as systems of di–erent administration share a single sign-on environment and access control lists; Fault tolerance, due to replication in local and remote storage systems; and Persistence, since data can be migrated to new local supports without a–ecting availability [11]. Figure 1 represents the federation model of organization of data grids.

2.3 iRODS

The iRODS[1] system is an open-source storage solution for data grids based on distributed client-server architecture. A database in a central repository, called iCAT, is used to maintain, among other things, the information about the nodes in the Grid, the state of data and its attributes, and information about users. A rule system is used to enforce and execute adaptive rules. This system belongs to the class of adaptive middleware systems, since it allows users to alter software functionalities without any recompilation. Figure 2 shows the UML deployment diagram of iRODS. Note that the iCAT database only resides in the central node and many other nodes can be connected to the central node.

iRODS uses the storage provided by the local file system, creating a virtual file system on top of it. That virtualization creates infrastructural independence, since logical names are given to files, users and resources. Management policies are mapped into rules that invoke and control operations (micro-services) on remote storage media. Rules can be used for access control, to access another grid system, etc. Middleware functions can be extended by composing new rules and policies.

The federation of multiple iRODS data grids is also a feature. Through the federation mechanism, an independent iRODS grid (i.e., an iCAT-enabled node and possibly zero or more non-iCAT servers connected in a grid) can interoperate with other independent iRODS grids. Each federated grid is called a *zone*. In

[1] https://www.irods.org

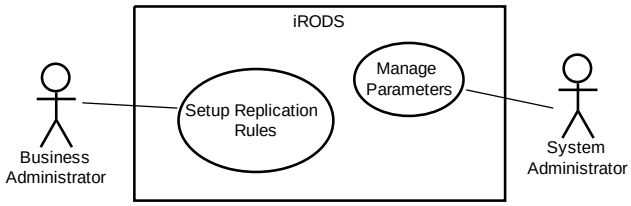


Figure 3 – Use case diagram of the proposal

a federated schema, a user with appropriate permissions can access objects stored in any iRODS node, belonging to any of the federated grids. iRODS also supports the existence of multiple federations.

2.4 Data Grids and Digital Preservation

Several publications and projects addressed the potential usage of data grids in digital preservation. The *InterPARES 2*[2] project studied the usage of data grid technologies for the building of preservation environments [12]. The conclusions were that many of the characteristics that were required to implement, manage and evolve a persistent archive were already present in existing data grids [13]. Data grids are also described as being useful for managing technological obsolescence, due to the virtualization of the underlying storage technologies [10].

In [14], the increasing needs of traditional archives and libraries for the preservation of large quantities of data is pointed as a driver for the collaboration with science and engineering partners for the use of data grid infrastructure. In the same reference, several US initiatives dealing with preservation using data grids are described. Data grids are also suggested for preservation purposes in [15].

As for concrete preservation solutions using data grids, in [16] an implementation of a prototype grid-based digital library using gLite[3] gLibrary[4] is described. iRODS extensibility features are explored in [17] to implement digital curation strategies. The use of the Storage Resource Broker (SRB) data grid for the preservation of digital media has been reported in [18]. iRODS usage for preservation purposes is explored in the SHAMAN project [19], while the DILIGENT[5] project explored the use of gLite [20]. The iRODS data grid technology was analyzed from the point of view of threats and vulnerabilities in [7].

3. THE PROBLEM

As already referred, redundancy is an important means to withstand failures that might endanger data. According to [3], redundancy is required to be able to recover data from storage media faults, natural disasters, human operational errors, internal attacks, and external attacks. In that sense, it can be said that redundancy is a crucial feature in any digital preservation solution.

Data grids in general offer the particularity of having, among other desired characteristics, a replication feature. Moreover, the fact that grids belonging to different institutions can engage in federations with other grids, while retaining full administrative control of their domains can be a useful feature in preservation scenarios, since additional storage space is obtained this way. This would allow a preservation system to take advantage of "borrowed" resources belonging to another administrative domain.

The iRODS data grid system supports both replication features and the creation of federations. However, it requires customization in order to take advantage of these features, namely through the micro-service mechanism. Micro-services are small and well-defined functions/procedures that execute a determined micro-level task. Users and administrators can chain micro-services in order to create macro-level functionalities (also called Actions). An example of a rule definition for an action is the following:

- *actionDef* | *condition* | *workflow-chain* | *recovery-chain*

The *actionDef* corresponds to the identifier of the rule. The *condition* field specifies a condition that must be met in order that the *workflow-chain* - a chain of micro-services - can be executed. Conditions can be one or more logical expressions. A workflow chain can be composed of several micro-services or actions (other rules), separated by "##" characters. The *recovery-chain* specifies a chain of recovery micro-services chain that will be executed in case something fails on the execution of the workflow-chain.

The composition of micro-services is not straightforward. iRODS already features a plethora of micro-services providing generic operations (for instance, a replication micro-service is already provided with iRODS). However, if one is looking for more specialized micro-services, programming skills are required. Furthermore, the composition of rules with the objective of implementing different kinds of data processing can also be cumbersome since it requires the learning of the syntax and direct editing of iRODS rule database. This requires that the person administrating the preservation system possesses strong technical skills which might be a barrier to the widespread adoption of this type of system as a digital preservation solution.

In addition to this, the federation feature of iRODS only allows a limited control of the resources and data of remote federated grids, due to each grid having its own administrative domain. For instance, access to data in a remote grid is possible for an authenticated user, but writing new data or updating existing data is a limited feature, requiring some tweaking.

4. THE PROPOSAL

In this section we describe our proposal, which is composed of an interface for the easy composition of replication rules, a compiler which transforms the composed rules into iRODS rules, a replication service which enforces the replication rules on the ingest of files, and an audit service which maintains the number of replicas. The replication and the audit services will have to run on each federated grid.

4.1 The Composition of Replication Rules

We can consider that we have two kinds of abstract actors: a Business Administrator, which is responsible for the creation and enforcement of replication rules and might not have strong technical skills, and a System Administrator, which is responsible for the administration and maintenance of the technical aspects of the system, and thus of replication.

[2] http://www.interpares.org/ip2/ip2_index.cfm

[3] http://glite.cern.ch/

[4] https://glibrary.ct.infn.it/glibrary_new/index.php

[5] http://diligent.ercim.eu/

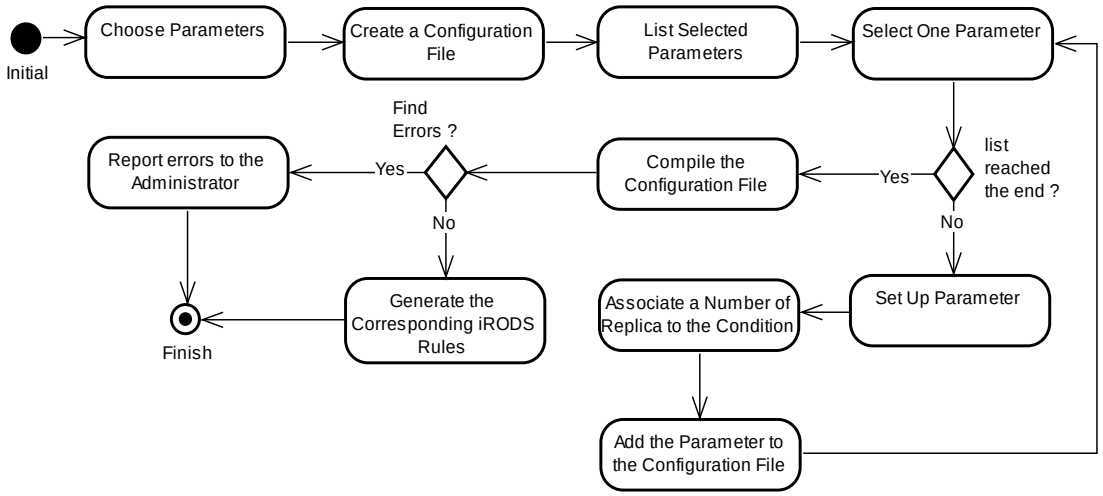


Figure 4 – Sequence diagram of the *Setup Replication Rules*

We propose an easy to use interface for managing replication rules so that business administrators without any specific technical skills can manage replication rules. That interface would support two use cases: *Setup Replication Rules* and *Manage Parameters*, which are represented in Figure 3.

In the *Setup Replication Rules* use case, the business administrator is able to create, in an easy way, specific rules based on determined parameters (e.g., file format, file size, user identity, etc.), in order to determine the minimum number of replicas to maintain of a file. The Parameters are listed as check boxes. The business administrator can select a list of parameters to be setup and then proceed to the configuration. When doing the configuration, for each parameter the business administrator can specify one condition (e.g., equal, different from, etc.) and enter a minimum number of replicas to be associated in case the condition is verified. The respectively configuration will be added to a configuration file. When finished, the configuration file is compiled and searched for any syntax errors. If no error is found, iRODS rules are generated and added to the rule base. Otherwise, the system will show the resultant errors. Figure 4 depicts the workflow sequence of this use case.

In the *Manage Parameters* use case, the system administrator can manage which parameters the business administrator can select and customize. The system can interpret certain parameters. The system administrator can choose from a previous list which parameters will be available to the business administrator. The system administrator can add or remove parameters to the actual list (the one that the business administrator can see), and apply the changes.

4.2 The Replication Service

Upon ingestion of a new file into the local iRODS deployment, the replication service checks if any of the rules configured by the business administrator apply to the file and, according to the rules, computes the number N of replicas to maintain of that file. That number is associated to the file through its metadata.

After that, the service checks the number of different federated grids. If the number of replicas is bigger or equal than the number of federated grids, a list of all federated grids is compiled. Otherwise, the first N federated grids listed are compiled into the list. Based on the list of federated grids available and on the number N of replicas, the number of replicas to be stored on the local iRODS deployment and on the remote federated grids is computed. The number of replicas to be stored in each zone, local or remote, is the integer division of N by the number of Zones. The remaining number of replicas until filling N is stored in the local Zone. The number of remote replicas R to be created is then registered and associated to the file metadata.

Then, one by one, each Zone contained in the federation list, will be used to create and store the replicas. If the local grid deployment is selected, the number of effectively created replicas L is registered and associated to the file metadata. If it is the case of a remote Zone, the file is copied and the number of remote replicas R that file should have is associated to that copied file. The file is not directly replicated, since the federation configuration does not allow the direct replication of files to a remote Zone. The file has to be copied and the replication has to be executed by the audit service running in the remote grid, which we will explain in the next section. Also, when copying a file to a remote zone, the associated metadata is not copied, hence the association of the desired number of replicas R after the copy. A reference to the file copied to the remote zone is also maintained. Figure 5 depicts the activity diagram of the replication service.

4.3 The Replica Audit Service

The replica audit service functions at two different levels. The first level audits the number of replicas stored in the local zone and the replicas of copies of local files stored in remote zones. The second level audits the number of replicas in the local zone which are owned by other zones (in other words, files which have been copied to the local zone from a remote zone).

Concerning the first level, a list of files contained in the local zone and owned by the local zone is compiled. Then, one by one, each file is selected and list of zones containing a replica of that file is compiled and, for each zone of the list, the number of replicas effectively stored in the selected zone is determined. When the list of zones is fully processed, if the total number of existing replicas is smaller than the replica number N contained in the file metadata, the number of necessary replicas in order to get N replicas is calculated. Then, a list containing all the zones where a copy of the file exists is compiled, and the number of replicas is increased to be as close as possible to L, in case of a

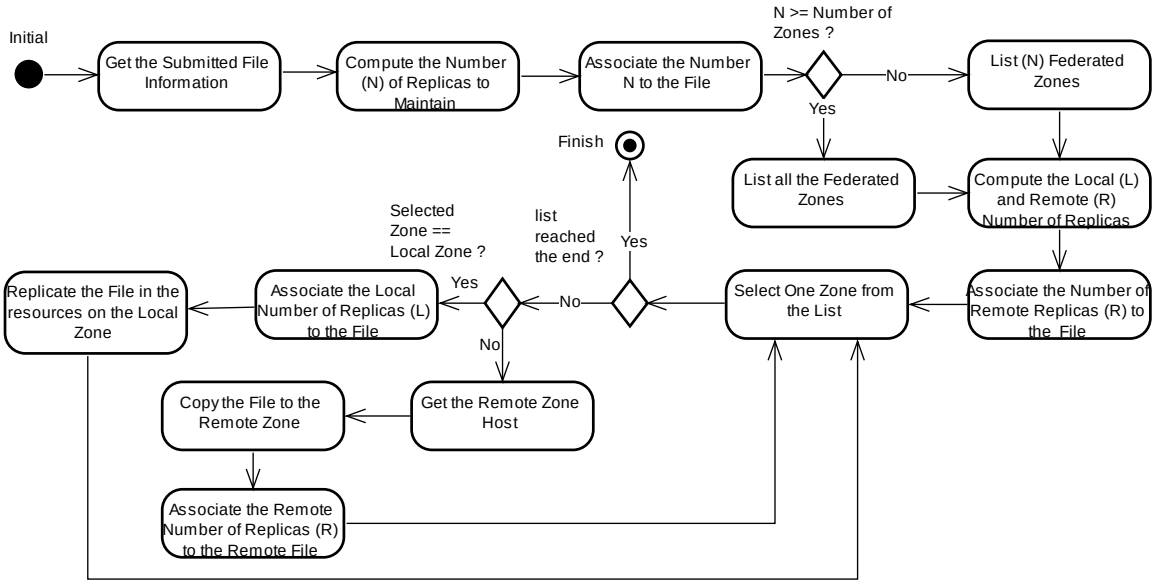


Figure 5 – Activity diagram of the *Replication Service*

local zone, or to R in case of a remote zone. After all files have been verified, an audit report of the status of each file is compiled and sent to the system administrator, so that, in case of need, he can take informed decisions. Figure 6 depicts the activity diagram of the first level of the auditing service.

The second level of audit begins with the creation of a list of files contained in the local zone which are owned by remote zones. Then, one by one, each file is selected and the number of replicas that the file should have (R) is retrieved. The number of accessible replicas is retrieved and is compared with the desired number of replicas. If the number of accessible replicas is smaller than R, than the number of replicas needed in order to reach R is calculated. If the number of existing resources is smaller than the number of replicas needed to reach R, a list of all the resources that do not contain a replica is retrieved. If the list is empty, then the issue is registered for the audit report and another file is selected for verification. Otherwise, files are replicated throughout available resources. After all files have been verified, an audit report is compiled and sent to the system administrator, so that, in case of need (e.g., add more storage resources) he can take informed measures. Figure 7 depicts the activity diagram of the second level of auditing.

4.4 Implementation

The interface for the composition of replication rules was implemented as a website, using HTML and PHP. The user is guided through the configuration of the rules. Currently supported replication parameters are *file extension*, *file name*, *file size*, *user*, *submission date*, and *resource name*. When the composition of rules is finished, an xml configuration file is generated and is directly processed by a compiler.

The compiler is written in C. For each replication parameter it should verify the syntax and generate the corresponding iRODS rules. The compiler output is a set of iRODS rules as an iRODS rule base file so it can be included in the iRODS installation.

Both *Replication* and *Audit* Services are implemented using the rule mechanism and workflow capabilities provided by iRODS. The services are defined as a set of actions within a rule. The actions are composed by a set of micro-services. For the development of those micro-services we used the C language API provided by iRODS.

The *Replication* Service is triggered by a file submission. To access the file information we use available iRODS session variables. We send this information as input to the rules previously generated by the compiler. The execution of the rule results in a minimum number of replicas. This number is then associated to the file as a metadata attribute, using a micro-service already packed with iRODS.

When using resources located in remote zones for replication, the file has to be copied to the remote zone and the number of copies to maintain is associated with the file metadata. The remote zone then takes care of the replication. While we had to develop a set of micro-services to perform some of the operations involved, we also used micro-services already included in the installation.

The Audit Service is composed by two iRODS rules, one for auditing files owned by the local zone, and the other for auditing the files owned by remote zones. Again, some micro-services were developed to specifically for this purpose. Other micro-services were already packed with iRODS, such as the case of *msiSendMail*, which is used to send the audit report to the system administrator.

5. CONCLUSIONS AND FUTURE WORK

Data grids are systems which possess characteristics which are highly desirable for digital preservation, such as replication. Replication makes possible the adoption of a data redundancy strategy which is crucial to withstand failures. In addition, the federation configuration model present is data grids such as iRODS allows the interoperability between independent data grid installations, thus making possible the sharing of resources.

This paper presented a proposal based on the customization of the iRODS platform to be able to take advantage of its replication and federation features. That proposal was implemented with basis on the rule engine mechanism which allows the creation of rules that

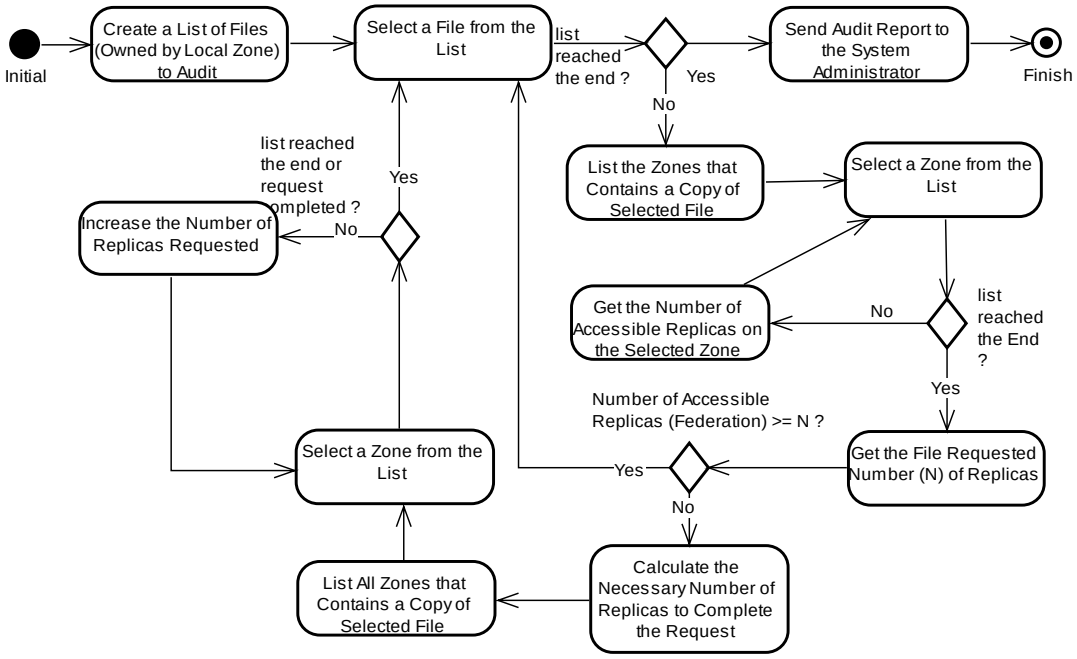


Figure 6 – Activity diagram of the *Audit of Locally Owned Files*

orchestrate micro-services. Through the creation of specialized micro-services, it was possible to create a complex replication service which takes advantage of the federation configuration, using the spare resources of federated grids to create replicas, thus using geographic distance and independent administration to lower the risk of losing data.

In addition to the replication service, an audit service was created, using the same mechanisms, which audits replicas at two levels: at the level of the locally-owned data, in which the data owned by a local grid, stored locally or remotely, can be audited; an at the level of remotely-owned data, in which the local grid audits data owned by remote federated grids, but stored locally.

Besides the implemented services, an interface for the composition of replication rules was also described. The use of a user-friendly interface would allow business administrators, with little technical knowledge, to define the rules applicable to the data. The kinds of rules that a business administrator can define have to be determined by the system administrator, more knowledgeable of technical aspects. The rules defined in the interface are then compiled into iRODS rules and included in the rules database.

Future work will focus on the validation of the proposed solution in the context of project SHAMAN[6] and TIMBUS[7]. Both projects address scenarios where the usage of data grids for preserving data assumes major relevance.

6. ACKNOWLEDGMENTS

This work was supported by FCT (INESC-ID multi-annual funding) through the PIDDAC Program funds and by the projects SHAMAN and TIMBUS, funded under FP7 of the EU under contract 216736 and 269940, respectively.

[6] http://www.shaman-ip.eu

[7] http://timbusproject.net/

7. REFERENCES

[1] IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries," New York, 1990.

[2] Miles, S., Wong, S. C., Fang, W., Groth, P., Zauner, K.-P. and Moreau, L. 2007. Provenance-based validation of e-science experiments," *Web Semant.*5 (Mar. 2007), 28–38.

[3] Barateiro, J., Antunes, G., Freitas, F. and Borbinha, J. 2010. Designing Digital Preservation Solutions: A Risk Management-Based Approach. *The International Journal of Digital Curation*. 1, 5 (Jun. 2010), 4-17.

[4] Johnston W. E. 2002. Computational and Data Grids in Large-scale Science and Engineering. *Future Gener. of Comput. Syst*. 18, 8, 1085-1100

[5] Venugopal, S., Buyya, R., and Ramamohanarao, K. A. 2006. Taxonomy of data grids for distributed data sharing, management, and processing. *ACM Computing Surveys*. 38,1, 1–53.

[6] Rajasekar, A., Wan, M., Moore, R. and Schroeder, W. 2006. A prototype rule-based distributed data management system. In *HPDC workshop on Next Generation Distributed Data Management* (Paris, France).

[7] Barateiro, J., Antunes, G., Cabral, M., Borbinha, J. and Rodrigues, R. 2008. Using a GRID for digital preservation. In *Proceeding of the International Conference on Asian Pacific Digital Libraries* (Bali, Indonesia).

[8] Foster, I. 2002. What is a grid? A three point checklist. *Grid Today*. 1(6).

[9] Chervenak, A., Foster, I., Kesselman, C., Salisbury, C., and Tuecke, S. 2000. The data grid: Towards an architecture for

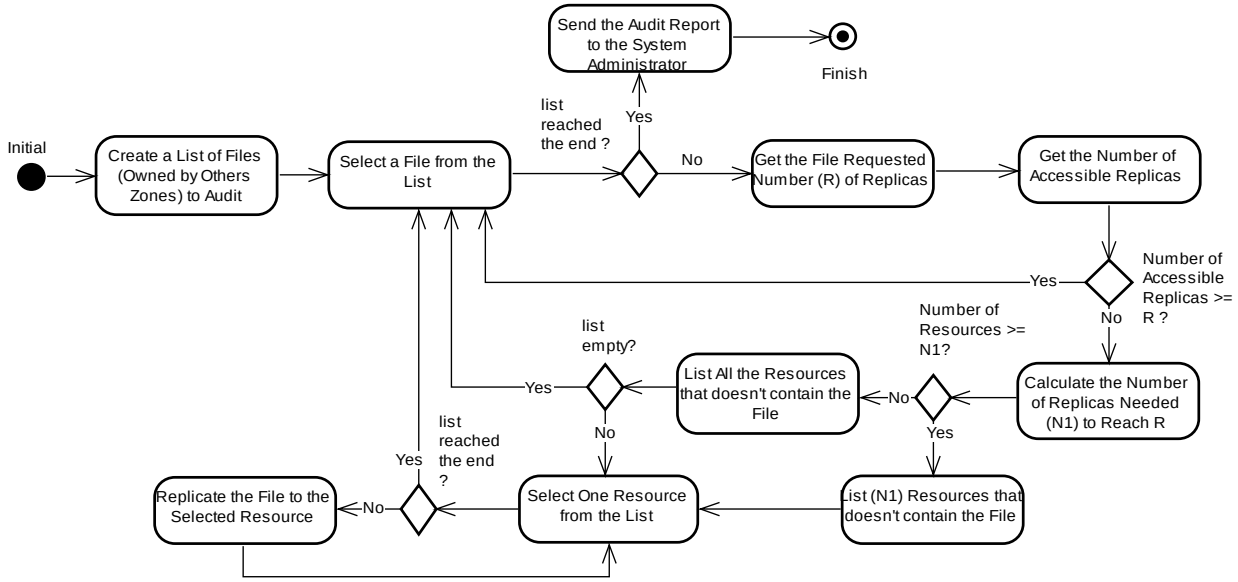


Figure 7 - Activity diagram of the *Audit of Remotely Owned Files*

the distributed management and analysis of large scientific datasets. *Journal of Network and Computer Applications.* 23, 187–200.

[10] Moore, R. 2005. Building preservation environments with data grid technology. *American Archivist*, 69, 1, 139–158.

[11] Rajasekar, A., Wan, M., and Moore, R. 2002. MySRB & SRB - components of a data grid. In *The 11th International Symposium on High Performance Distributed Computing* (Edinburgh, Scotland). (HPDC-11).

[12] Duranti, L. 2005. The Long-Term Preservation of Accurate and Authentic Digital Data: The InterPARES Project. *Data Science Journal*. 4, 25 (Oct. 2005), 106-118.

[13] Moore, R. 2007. General Study 01 Final Report: Building Preservation Environments with Data Grid Technology. InterPARES 2 Project.

[14] Jordan, C., Kozbia, A., Minor, D. and McDonald, R. H. 2008. Encouraging Cyberinfrastructure Collaboration for Digital Preservation. In *Proc. iPRES2008* (London, UK).

[15] Gao, J., Li, Z., Wang, X. and Zhu, C. 2009. Research on Grid Storage Technology and its Application in Digital Library. In *the 2nd International Symposium in Knowledge Acquisition and Modeling* (Wuhan, China).

[16] Calanducci, A. S., Barbera, R., Cedillo, J. S., De Fillipo, A., Saso, M., Iannizzotto, S., De Mattia, F. and Vicinanza D. 2009. Data Grids for Conservation of Cultural Inheritance. In *Proc. DaGreS '09* (Ischia, Italy).

[17] Hedges, M., Hasan, A., Blanke, T. 2007. Management and Preservation of Research Data with iRODS. In *Proc. of CIMS '07* (Lisbon, Portugal).

[18] Chien-Yi Hou, Altintas, I., Jaeger-Frank, E., Gilbert, L., Moore, R., Rajasekar, A. and Marciano, R. 2006. A scientific workflow solution to the archiving of digital media. In *Workshop on Workflows in Support of Large-Scale Science* (WORKS '06)

[19] Innocenti, P., Ross, S., Maceiviciute, E., Wilson, T., Ludwig, J. and Pempe, W. 2009. Assessing Digital Preservation Frameworks: the Approach of the SHAMAN Project. In *Proc. of MEDES '09* (Lyon, France).

[20] Candela, L., Akal, F., Avancini, H., Castelli, D., Fusco, L., Guidetti, V., Langguth, C., Manzi, A., Pagano, P., Schuldt, H., Simi, M., Springmann, M. and Voicu, L. 2007. DILIGENT: integrating digital library and Grid technologies for a new Earth observation research infrastructure. In *Int. J. Digit. Libr.* 7, 59-80.

Using Automated Dependency Analysis To Generate Representation Information

Andrew N. Jackson
The British Library
Boston Spa, Wetherby
West Yorkshire, LS23 7BQ, UK
Andrew.Jackson@bl.uk

ABSTRACT

To preserve access to digital content, we must preserve the representation information that captures the intended interpretation of the data. In particular, we must be able to capture performance dependency requirements, i.e. to identify the other resources that are required in order for the intended interpretation to be constructed successfully. Critically, we must identify the digital objects that are only referenced in the source data, but are embedded in the performance, such as fonts. This paper describes a new technique for analysing the dynamic dependencies of digital media, focussing on analysing the process that underlies the performance, rather than parsing and deconstructing the source data. This allows the results of format-specific characterisation tools to be verified independently, and facilitates the generation of representation information for any digital media format, even when no suitable characterisation tool exists.

Categories and Subject Descriptors

D.2.8 [**Software Engineering**]: Metrics—*complexity measures, process metrics*; H.m [**Information Systems**]: Miscellaneous

General Terms

Algorithms, Measurement, Verification

1. INTRODUCTION

When attempting to preserve access to digital media, keeping the bitstreams is not sufficient - we must also preserve information on how the bits should be interpreted. This need is widely recognised, and this data is referred to as Representation Information (RI) by the Open Archival Information System (OAIS) reference model [4]. The reference model also recognises that software can provide valuable RI, expecially when the source code is included. However, software is not the only dynamic dependency that must be captured in order to preserve access. The interpretation of a digital object may inherit further information from the technical environment as the performance proceeds, such as passwords or licenses for encrypted resources, default colour spaces, page dimensions or other rendering parameters and, critically, other digital objects that the rendering requires. This last case can include linked items that, while only referenced in the original data, are included directly in the performance. In the context of hypertext, the term 'transclusion' has been coined to describe this class of included resource [5].

The classic example of a transcluded resource is that of fonts. Many document formats (PDF, DOC, etc.) only reference the fonts that should be used to render the content via a simple name (e.g. 'Symbol'), and the confusion and damage that these potentially ambiguous references can cause has been well documented [1]. Indeed, this is precisely why the PDF/A standard [2] requires that all fonts, even the so-called 'Postscript Standard Fonts' (e.g. Helvetica, Times, etc.), should be embedded directly in archival documents instead of merely referenced. Similarly, beyond fonts, there are a wide range of local or networked resources that may be transcluded, such as media files and plug-ins displayed in web pages, documents and presentations, or XSD Schema referenced from XML. We must be able to identify these different kinds of transcluded resources, so that we can either include them as explicit RI or embed them directly in the target item (as the PDF/A standard dictates for fonts).

Traditionally, this kind of dependency analysis has been approached using normal characterisation techniques. Software capable of parsing a particular format of interest is written (or re-used and modified) to extract the data that indicates which external dependencies may be required. Clearly, creating this type of software requires a very detailed understanding of the particular data format, and this demands that a significant amount of effort be expended for each format of interest. Worse still, in many cases, direct deconstruction of the bitstream(s) is not sufficient because the intended interpretation deliberately depends on information held only in the wider technical environment, i.e. the reference to the external dependency is implicit and cannot be drawn from the data.

This paper outlines a complementary approach, developed as part of the SCAPE project[1], which shifts the focus from the data held in the digital file(s) to the process that underlies the performance. Instead of examining the bytes, we use

[1] http://www.scape-project.eu/

iPRES2011, Nov. 1–4, 2011, Singapore.

the appropriate rendering software to walk-through or simulate the required performance. During this process we trace certain operating system operations to determine which resources are being used, and use this to build a detailed map of the additional RI required for the performance, including all transcluded resources. Critically, this method does not require a detailed understanding of file format, and so can be used to determine the dependencies of a wide range of media without the significant up-front investment that developing a specialised characterisation tool requires.

2. METHOD

Most modern CPUs can run under at least two operating modes: 'privileged' mode and 'user' mode. Code running in privileged mode has full access to all resources and devices, whereas code running in user mode has somewhat limited access. This architecture means only highly-trusted code has direct access to sensitive resources, and so attempts to ensure that any badly-written code cannot bring the whole system to a halt, or damage data or devices by misusing them. However, code running in user space must be able to pass requests to devices, e.g. when saving a file to disk, and so a bridge must be built between the user and the protected modes. It is the responsibility of the operating system kernel to manage this divide. To this end, the kernel provides a library of system calls that implement the protected mode actions that the user code needs.

Most operating systems come with software that allows these 'system calls' to be tracked and reported during execution, thus allowing any file system request to be noted and stored without interfering significantly with the execution process itself[2]. The precise details required to implement this tracing approach therefore depend only upon the platform, i.e. upon the operating system kernel and the software available for monitoring processes running on that kernel.

This monitoring technique allows all file-system resources that are 'touched' during the execution of any process to be identified, and can distinguish between files being read and files being written to. This includes software dependencies, both directly linked to the original software and executed by it, as well as media resources.

Of course, this means the list of files we recover includes those needed to simply run the software as well as those specific to a particular digital media file. Where this causes confusion, we can separate the two cases by, for example, running the process twice, once without the input file and once with, and comparing the results. Alternatively, we can first load the software alone, with no document, and then start monitoring that running process just before we ask it to load a particular file. The resources used by that process can then be analysed from the time the input file was loaded, as any additional resource requirements must occur in the wake of that event.

2.1 Debian Linux

On Linux, we can make use of the standard system call tracer 'strace', which is a debugging tool capable of printing out a trace of all the system calls made by another process or program[3]. This tool can be compiled on any operating system based on a reasonably recent Linux kernel, and is available as a standard package on many distributions. In this work, we used Debian Linux 6.0.2 and the Debian strace package[4]. For example, monitoring a process that opens a Type 1 Postscript (PFB) font file creates a trace log that looks like this:

```
5336 open("/usr/share/fonts/type1/gsfonts/
   n019004l.pfb", O_RDONLY) = 4
5336 read(4, "\200\1\f\5\0\0%!PS-
   AdobeFont-1.0: Nimbus"..., 4096) = 4096
      ...more read calls...
5336 read(4, "", 4096) = 0
5336 close(4) = 0
```

Access to software can also be tracked, as direct dependencies like dynamic linked libraries (e.g. '/usr/lib/libMagickCore.so.3') appear in the system trace in exactly the same way as any other required resource. As well as library calls, a process may launch secondary 'child' processes, and as launching a process also requires privileged access, these events be tracked in much the same way (via the 'fork' or 'execve' system calls). The strace program can be instructed to track these child processes, and helpfully reports a brief summary of the command-line arguments that we passed when a new process was launched.

2.2 Mac OS X

On OS X (and also Solaris, FreeBSD and some others) we can use the DTrace tool from Sun/Oracle[5]. This is similar in principle to strace, but is capable of tracking any and all function calls during execution (not just system calls at the kernel level). DTrace is a very powerful and complex tool, and configuring it for our purposes would be a fairly time-consuming activity. Fortunately, DTrace comes with a tool called 'dtruss', which pre-configures DTrace to provide essentially the same monitoring capability as the strace tool. The OS X kernel calls have slightly different names, the format of the log file is slightly different, and the OS X version of DTrace is not able to log the arguments passed to child processes, but these minor differences do not prevent the dependency analysis from working.

2.3 Windows

Windows represents the primary platform for consumption of a wide range of digital media, but unfortunately (despite the maturity of the operating system) it was not possible to find a utility capable of reliably assessing file usage. The 'SysInternals Suite'[6] has some utilities that can identify which files a process is currently accessing (such as Process Explorer or Handle) and similar utilities (ProcessActivityView, OpenedFilesView) have been published by a

[2]The tracing does slow the execution down slightly, mostly due to the I/O overhead of writing the trace out to disk, but the process is otherwise unaffected.

[3]http://sourceforge.net/projects/strace/
[4]http://packages.debian.org/stable/strace
[5]http://opensolaris.org/os/community/dtrace/
[6]http://technet.microsoft.com/en-gb/sysinternals/bb842062

third-party called Nirsoft[7]. These proved difficult to invoke as automated processes, and even when this was successful, the results proved unreliable. Each time the process was traced, a slightly different set of files would be reported, and files opened for only brief times did not appear at all. Sometimes, even the source file itself did not appear in the list, proving that important file events were being missed. This behaviour suggests that these programs were rapidly sampling the usage of file resources, rather than monitoring them continuously.

An alternative tool called StraceNT[8] provides a more promising approach, as it can explicitly intercept system calls and so is capable of performing the continuous resource monitoring we need. However, in its current state it is difficult to configure and, critically, only reports the name of the library call, not the values of the arguments. This means that although it can be used to tell if a file was opened, it does not log the file name and so the resources cannot be identified. However, the tool is open source, so might provide a useful basis for future work.

One limited alternative on Windows is to use the Cygwin UNIX-like environment instead of using Windows tools directly. Cygwin comes with its own strace utility, and this has functionality very similar to Linux strace. Unfortunately, this only works for applications built on top of the Cygwin pseudo-kernel (e.g. the Cygwin ImageMagick package). Running Windows software from Cygwin reports nothing useful, as the file system calls are not being handled by the Cygwin pseudo-kernel.

3. RESULTS

In this initial investigation, we looked at two example files, covering two different media formats that support transcluded resources: a PDF document and a PowerPoint presentation.

3.1 PDF Font Dependencies

The fonts required to render the PDF test file (the 'ANSI/NISO Z39.87 - Data Dictionary - Technical Metadata for Digital Still Images' standards document [3]) were first established by using a commonly available tool, pdffonts[9], which is designed to parse PDF files and look for font dependencies. This indicated that the document used six fonts, one of which was embedded (see Table 1 for details).

The same document was rendered via three different pieces of software, stepping through each page in turn either manually (for Adobe Reader or Apple Preview) or automatically. The automated approach simulated the true rendering process by rendering each page of the PDF to a separate image via the ImageMagick[10] conversion commmand 'convert input.pdf output.jpg'. This creates a sequence of numbered JPG images called 'output-###.jpg', one for each page.

All system calls were traced during these rendering processes, and the files that the process opened and read were collated. These lists were then further examined to pick out all of the dependent media files - in this case, fonts. The reconstructed font mappings are shown in Table 1.

The two manual renderings on OS X gave completely identical results, with each font declaration being matched to the appropriate Microsoft TrueType font. The manual rendering via Adobe Reader on Debian was more complex. The process required three font files, but comparing the 'no-file' case with the 'file' case showed that the first two (DejaVuSans and DejaVuSans-Bold) were involved only in rendering the user interface, and not the document itself. The third file, 'ZX_____.PFB', was supplied with the Adobe Reader package and upon inspection was found to be a Type 1 Postscript Multiple Master font called 'Adobe Sans MM', which contains all the variants of a typeface that Adobe Reader uses to render standard or missing fonts. Adobe have presumably taken this approach in order to ensure the standard Postscript fonts are rendered consistently across platforms, without depending on any external software packages that are beyond their control.

Although the precise details and naming conventions differed between the platforms, each of the ImageMagick simulated renderings pulled in the essentially the same set of Type 1 PostScript files, which are the open source (GPL-compatible license) versions of the Adobe standard fonts. This is not immediately apparent due to the different naming conventions using on different installations, but manual inspection quickly determined that, for example, NimbusSanL-Bold and n019004l.pfb were essentially the same font, but from different versions of the gsfonts package. The information in the system trace log made it easy to determine how ImageMagick was invoking GhostScript, and to track down the font mapping tables that GhostScript was using to map the PDF font names into the available fonts.

Interestingly, as well as revealing that these apparently identical performances depend on different versions of different files in two different formats (TrueType or Type 1 Postscript fonts), the results also show that while Apple Preview and ImageMagick indicate that Times New Roman is a required font (in agreement with the pdffonts results) this font is not actually brought in during the Adobe Reader rendering processes. A detailed examination of the source document revealed that while Times New Roman is declared as a font dependency on one page of the document, this appears to be an artefact inherited from an older version of the document, as none of the text displayed on the page is actually rendered in that font.

3.2 PowerPoint with Linked Media

A simple PowerPoint presentation was created in Microsoft PowerPoint for Mac 2011 (version 14.1.2), containing some text and a single image. When placing the image, PowerPoint was instructed to only refer to the external file, and not embed it, simulating the default behaviour when including large media files. The rendering process was then performed manually, looking through the presentation while tracing the system calls. As well as picking up all the font dependencies, the fact that the image was being loaded from an external location could also be detected easily.

The presentation was then closed, and the referenced image

[7]http://www.nirsoft.net/
[8]https://github.com/ipankajg/ihpublic/
[9]Part of Xpdf: http://foolabs.com/xpdf/
[10]http://www.imagemagick.org/

Tool	Operating System	List of Fonts
pdffonts 3.02	OS X 10.7	Arial-BoldMT. ArialMT, Arial-ItalicMT, Arial-BoldItalicMT TimesNewRomanPSMT, BBNPHD+SymbolMT (embedded)
Apple Preview 5.5	OS X 10.7	/Library/Fonts/Microsoft/... Arial Bold.ttf, Arial.ttf, Arial Italic.ttf, Arial Bold Italic.ttf, Times New Roman.ttf
Adobe Reader X (10.1.0)	OS X 10.7	/Library/Fonts/Microsoft/... Arial Bold.ttf, Arial.ttf, Arial Italic.ttf, Arial Bold Italic.ttf
Adobe Reader 9.4.2	Debian Linux 6.0.2	/usr/share/fonts/truetype/ttf-dejavu/... DejaVuSans.ttf, DejaVuSans-Bold.ttf /opt/Adobe/Reader9/Resource/Font/ZX_____.PFB
ImageMagick 6.7.1	OS X 10.7 via MacPorts	/opt/local/share/ghostscript/9.02/Resource/Font/... NimbusSanL-Bold, NimbusSanL-Regu, NimbusSanL-ReguItal, NimbusSanL-BoldItal, NimbusRomNo9L-Regu
ImageMagick 6.6.0	Debian Linux 6.0.2	/usr/share/fonts/type1/gsfonts/... n019004l.pfb, n019003l.pfb, n019023l.pfb, n019024l.pfb, n021003l.pfb
ImageMagick 6.4.0	Cygwin on WinXP	/usr/share/ghostscript/fonts/... n019004l.pfb, n019003l.pfb, n019023l.pfb, n019024l.pfb, n021003l.pfb

Table 1: Font dependencies of a specific PDF document, as determined via a range of tools.

was deleted. When re-opening the presentation, the system call trace revealed that PowerPoint was hunting for the missing file, guessing a number of locations based on the original absolute pathname. This approach can therefore be used to spot missing media referenced by PowerPoint presentations.

4. CONCLUSIONS

Process monitoring and system call tracing is a valuable analysis technique, complementary to the more usual format-oriented approach. It enables us to perform detailed quality assurance of existing characterisation tools, using a completely independent approach to validate the identification of the resources required to render a digital object. Furthermore, because the tracing process depends only on standard system functionality, and not on the particular software in question, it can work for all types of digital media without developing software for each format. As the PowerPoint example shows, the only requirement for performing this analysis is the provision of suitable rendering software.

Before using this approach in a production setting, it will be necessary to test it over a wider range of documents and types of transclusion, e.g. embedded XML Schema. In particular, the monitoring should be extended to track network requests for resources as well as local file or software calls. Although all network activity is visible via kernel system calls, the raw socket data is at such a low level that it is extremely difficult to analyse. Fortunately, tools like netstat[11] and WireShark[12] have been designed to solve precisely this problem, and could be deployed alongside system call tracing to supply the necessary intelligence on network protocols. Beyond widening the range of resources, extending this approach to the Windows platform would be highly desirable. The current lack of a suitable call tracing tool is quite unfortunate, and means that this approach cannot be applied to software that only runs on Windows. Hopefully, StraceNT can provide a way forward.

[11]http://en.wikipedia.org/wiki/Netstat
[12]http://www.wireshark.org/

Beyond the direct resource dependencies outlined here, this approach could be combined with knowledge of the platform package management system in order to build an even richer model of the representation information network a digital object requires. For example, Debian has a rigorous package management processes, and by looking up which packages provide the files implicated in the rendering, we can validate not only the required binary software packages, but also determine the location of the underlying open source software, and even the identities of the developers and other individuals involved. This allows very rich RI to be generated in an automated fashion. Furthermore, as the Debian package management infrastructure also tracks the development and discontinuation of the various software packages, this information could be leveraged to help build a semi-automatic preservation watch system.

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] G. Brown and K. Woods. Born Broken : Fonts and Information Loss in Legacy Digital Documents. *International Journal of Digital Curation*, 6(1):5–19, 2011.

[2] International Standardization Organization. ISO 19005-1:2005 Document management – Electronic document file format for long-term preservation – Part 1: Use of PDF 1.4 (PDF/A-1), 2005.

[3] National Information Standards Organization. ANSI/NISO Z39.87 - Data Dictionary - Technical Metadata for Digital Still Images, 2006.

[4] The Consultative Committee for Space Data. Reference Model For An Open Archival Information System (OAIS), 2009.

[5] Theodor Holm Nelson and Robert Adamson Smith. Back to the Future, 2007.

Cyberinfrastructure Supporting Evolving Data Collections

Maria Esteva
Texas Advanced Computing Center
maria@tacc.utexas.edu

Christopher Jordan
Texas Advanced Computing Center
ctjordan@tacc.utexas.edu

Tomislav Urban
Texas Advanced Computing Cen
turban@tacc.utexas.edu

David Walling
Texas Advanced Computing Center
walling@tacc.utexas.edu

ABSTRACT
The requirements to support large-scale and complex research collections are growing at an accelerated pace. Considering the continuous evolution of the collections, their increasing sizes, the technologies supporting them, and the importance of adequate data management to long-term preservation, a team at the Texas Advanced Computing Center (TACC) developed a cyberinfrastructure to aid researchers in the creation, management, and curation of collections throughout the research lifecycle processes and beyond for access and long term preservation. Collections are maintained on a petabyte-scale data applications facility, and consulting services are available to address data curation needs. In this environment, researchers have the flexibility to build their collections without having to deal with details such as systems administration and hardware migration planning. The cyberinfrastructure facilitates the development of sustainable collections and a seamless transition through data gathering and curation, large-scale analysis, and collections dissemination and preservation.

Categories and Subject Descriptors
H.3.2. **[Information Storage]:** Record Classification; H.3.7 [Digital Libraries]: Collection, Dissemination, Standards.

General Terms
Management, Design, Economics, Reliability

Keywords
Data management, preservation, storage architecture, metadata

1. INTRODUCTION
In the last 10 years there has been a noticeable impulse towards early implementation of data management strategies as a fundamental step towards preserving the digital products of research [1]. Culminating in funding agencies' requirements to include data management plans in the grant proposals [2], this development is driven by the recognized value of digital collections reuse for knowledge dissemination and data-driven discoveries. Currently, many academic libraries are implementing information services to help researchers craft their data management plans as well as providing guidance on how to deposit their collections for long-term preservation, often within their own institutional repositories (IR). While valuable, these services may fall short in addressing complex and large-scale collections' architectures, and in meeting the technical and curatorial needs that emerge during their development stages.

Research data collections are increasingly becoming complex systems, formed by diverse data objects included within layers of software and hardware that provide functionalities needed for analysis and interaction with the data. While research is conducted, raw data from experiments and observations are transformed during analysis, at the same time as new data is being gathered, incorporated into subsequent pipelines, curated, published, and archived. These juxtaposed processes are often difficult to document and some collections growth can be indefinite. Through the research lifecycle, researchers may use different applications to process their data and hardware to store it, resulting in dispersed and often incompatible datasets, which creates research bottlenecks and places data under risk [3]. As collections evolve, so do the standards and the expertise required to support them, which differ across and within science domains. In turn, the technologies on which these collections depend change at fast pace, while the cycles of funding to upgrade and maintain them are uneven. Without data management strategies and an adequate technical environment to facilitate these processes, the possibilities for long-term access and reuse of these collections are challenged. Thus, supporting the continuing and sustainable development of these *evolving collections* requires rethinking the infrastructure and service models.

In 2008 the Texas Advanced Computing Center (TACC) at the University of Texas at Austin (wwww.tacc.utexas.edu), established the Data Management and Collections group (DMC) to work with researchers in the Sciences, Engineering, Social Sciences, and Humanities fields in lifecycle management of research collections. Confronted with the diversity of data and requirements presented by research teams seeking solutions for their collections, we articulated the concept of evolving collections that allows mapping curatorial and technical tasks to the different research stages. Based on this concept, we developed a flexible storage facility and data management services as a cyberinfrastructure to meet the researchers' needs.

Developed in the context of a supercomputing center that enables researchers to conduct large computational tasks, the cyberinfrastructure is designed to seamlessly integrate processes from data gathering, to analysis, dissemination, and preservation. Within such an environment, and throughout the stages during which data is transformed into collections, research teams can conduct curatorial and technical tasks while users access the data. With maintenance and security of the infrastructure provided by TACC, researchers can remain focused on their research without having to deal with day-to-day systems operations. As a

iPRES2011, Nov. 1–4, 2011, Singapore.

contribution to the ongoing discussion in digital curation and preservation, in this paper we present our activities and resources illustrated with vignettes from the collections currently supported at TACC.

2. RELATED WORK

Two general models are predominant in data preservation: 1) the centralized model, in which a repository preserves a collection after it is finalized in standard, archival formats; and 2) the decentralized model, in which research teams curate and give access to their own data. Neither is particularly well suited to address the transformations and technical challenges data undergo through the research process. The centralized model cannot address the needs of ongoing research with an evolving collection, while the decentralized one often neglects data management strategies, as these may be too burdensome or technically challenging for the researchers to accomplish [3].

The evolving collections cyberinfrastructure operates in a middle area between these two models. It enables researchers to develop and archive collections in a continuum at any point in their research lifecycle, while they, other users or both perform data analysis and visualization tasks across storage and computing environments. Recently, projects have emerged to allow users to move their data across cloud storage providers and access it as needed, with the added value of including preservation services to assure data integrity and transactions transparency [5]. Like TACC's services, this model proposes maintaining the infrastructure for the users. Differently, it does not yet provide flexibility for researchers to build complex and unique collections architectures and functionalities in the cloud. Thus, as long as the workflows are not fully integrated in the cloud, researchers may have to maintain both local and cloud instances of their data and applications [5]. In contrast, we can provide the same interfaces and/or functionality as DuraCloud on top of our resources. Just as we run database and web servers, we could run DuraSpace and associated applications for users that want those interfaces.

While supercomputing centers are being considered in current data curation discussions [6], their role in data stewardship is not fully developed and their activities and potential are not well known to the community. This paper clarifies the expertise and resources available at TACC, and suggests a scalable model for collections preservation. The cyberinfrastructure has a post-custodial flair [4], as the researcher remains the curator and owner of the collection, while TACC provides the data facility and consulting for as long as it is agreed upon.

3. CYBERINFRASTRUCTURE

3.1 Team Expertise

The DMC group designs, builds, and maintains the data applications facilities, and consults with researchers in aspects of their collections, from creation to long-term preservation and access. Group members are specialized in software application development, Relational Database Management Systems (RDBMS), Geographical Information Systems, scientific data formats, metadata, large storage architecture, system administration, and digital archiving and long-term preservation.

3.2 Infrastructure

Lifecycle collections activities are centralized in a data applications facility called Corral, which provides different technical environments as collections building blocks that users may select according to the functionalities that they need. Corral, consists of 1.2 Petabytes of online disk and a number of servers providing high-performance storage for all types of digital data. It supports databases, web-based access, and other network protocols for storage and retrieval of data. A high-performance parallel file system based on Lustre is directly accessible from all of TACC's High Performance Computing (HPC) resources, enabling mathematical computation and visual analyses of petabyte-scale datasets. Corral's disk subsystem provides 6GB/sec of performance, and each of the web and file servers can move data from between 100MB/sec and 1GB/sec. While data can be moved through multiple servers and services at full speed simultaneously, the system is configured so that no one user or service can consume all the available performance.

For database collections, Corral provides flexibility in terms of the RDBMSs that users may choose from. DB server nodes running MySQL, PostgreSQL, and SQL Server are available. The DB nodes can be easily accessed at high bandwidth by web applications running on Corral's web server nodes. We also maintain open source domain specific databases such as ARK [7] and Specify [8], which support Archaeology and Natural History collections respectively. This infrastructure is useful to researchers from the Pecan Street Project (www.pecanstreetproject.org). The team needed a workflow for ingesting nightly dumps of the usage information of electrical devices from 100+ wired homes in Austin TX, into a database system on which analysis of energy usage can be conducted. A MySQL database is implemented in Corral to serve an ongoing data collection of over 1 billion records, with 5M+ new records added everyday. The size of the datasets and type of general query analysis conducted at this stage, led us to exploring OLAP and column oriented database services for enabling quick analysis, as well as evaluating solid state disks for increasing performance of the SQL queries conducted by the researchers.

Collections requiring long-term preservation are managed within iRODS [9]. Off site replication is done in Ranch, a Sun Microsystems StorageTeck Mass Storage tape system with a capacity of 10 PB, and geographical replication is accomplished through an agreement with Indiana University's Research Computing Division [10]. Close supervision, parts replacement contracts, and frequent schedule of upgrades are in place for maintaining the infrastructure. This model is based on TACC's experience managing systems to assure 24/7 services and data security,

Coupled with performance levels for reliable data transfer between storage and computing resources, the integrated infrastructure of Corral enables flexibility to implement different collection configurations and functionalities. The UT Center for Space Research (www.csr.utexas.edu) uses Corral to store very large sensor, satellite, aerial and radar datasets that they curate for dissemination purposes. Within three days of the 2010 earthquake in Haiti in collaboration with TACC, the repository/file system used for managing CSR data in Corral, was turned into a web repository for sharing data. This allowed CSR to access, organize, retrieve, and post the data required by the emergency operations in the region [11]. This type of quick repurposing allowed a multi-terabyte collection managed through one application, to instantly become accessible through a password-protected web application on another server.

3.3 Services

3.3.1 Collections Set-up

Users can request a storage allocation through TACC's user portal and select services to build, manage, and archive their collections.

Storage allocations are renewed on a yearly bases, including revisiting the services needed.

Significant consulting with group members takes place before the data is transferred to TACC. This allows planning data transfers, and deciding what technologies and configurations are needed for a particular collection. Examples of such work entail documenting the existing collection's architecture, guiding data inventories, analyzing data pipelines and improving aspects of their organization, and implementing metadata standards. To set up a collection, and depending on its architecture and lifecycle stage, DMC members manage access to the systems, install database servers, dependency libraries and webservers, and migrate data as needed. Users have access to their deployed web code, but are not burdened with systems administration tasks. A simple case of collection set up is the Oplontis archaeology project (www.oplontisproject.org). To facilitate remote access to a team of international researchers so they could input data and interpretations to a database, we moved an SQL database located at a restricted server to an SQL server on Corral. More complex projects require migrating data and code from commercial databases to ones that run on Corral, and integrating data and metadata from diverse legacy systems.

An ongoing effort is automating and generalizing services. For large image collections like the University of Alaska Museum's Herbarium (www.uaf.edu/museum/collections/herb), as the curators upload the raw files, processing scripts create image derivatives and OCR of labels. In turn, these scripts can be adapted for parallel processing of very large image collections from the UT Libraries in TACC's HPC resources. Rules based services for iRODS are also refactored to use across different collections. Once collections are fully functional and some services are automated, users require less specialized support. As they become their own collections managers, the tasks for our group are related to general data facility administration. When needed, users can submit tickets with requests for support via the TACC users support system.

3.3.2 Data Ingest and Retrieval

Data transfer to the system is achieved from various ingest tools, the selection of which depend on the needs of the users. For users conducting bulk submissions from their desktops to Corral, we developed TACCingest, to move large batches of files in a simple and reliable fashion. The tool was first tested so that Lawrence McFarland, a photography professor with a terabyte sized collection of ~3 GB images, can move large groups of images from his 100 hard-drives to safe storage. Command line and UI tools such as iDROP are also available to transfer data to iRODS, or to query its metadata catalog for data of interest. Data ingest activities may include automated metadata extraction, integrity checking and evaluation of file naming compliance.

The iPlant Collaborative (www.iplantcollaborative.org), which integrates data from standard plant genetic repositories as well as user submitted data, is illustrative of complex data transfers and retrieval due to the amount of users involved, and the functionalities that they request to use their data. Services involve developing applications to support large data ingest into iRODS, and a number of web interfaces to iRODS to make the data accessible to and from different analysis workflows. Provenance metadata is also collected and stored into iRODS using the same web API. In our configuration, digital objects may be accessed from a different workflow from which they originated, repurposed for analysis and or publication, and re-entered to the system as new objects.

3.3.3 Data Preservation and Integrity

Preservation services for the collections stored in iRODS include: rules to generate file checksums, automatic off-site and geographical replication, massive extraction of metadata using FITS [12] and encoded as Preservation Metadata (PREMIS), and finally registering the metadata in the iRODS catalogue.

Beyond basic bit level preservation and technical metadata gathering, we also address the domain scientists' conception of data preservation. In the case of archaeology collections, preservation is strongly associated with integrity, which involves maintaining the relationships between the objects found in a same context in the excavation. To assure that the archived data could render a representation of the site, The Institute of Classical Archaeology (www.utexas.edu/research/ica) selected to have two collection instances within Corral. A presentation instance resides on the ARK database and web site, which provides interactivity features and the possibility for users to study data objects in relation to their geospatial location and to the researchers' interpretations. The archival instance, stored in a hierarchical directory structure in iRODS, and replicated in Ranch, preserves contextual relationships between the raw images—and their versions—of the objects found on the excavation and their correspondent documentation, which is generated on the site and through the research lifecycle. These relationships are preserved through a context code recorded in the digital objects file names and in their metadata records, so that when one object is retrieved, all of the related objects are retrieved as well. In this way, if the ARK database ceases to be supported, the archival instance will serve to reconstruct the site.

3.3.4 Descriptive Metadata

When possible, to facilitate collections organization and avoid manual metadata entry, descriptive metadata is automatically extracted from the collections record-keeping system at ingest. This process requires the existence of an informative and regular file naming and or directory labeling across the collection. It also involves previous work mapping the descriptive data points to standard metadata schemas such as Dublin Core (DC) or Visual Resources Association (VRA) Core. The latter results from consulting with our team and training users on the required standards and practices. Implemented as an iRODS rule, a Jython script parses directory labels and file naming conventions as files are ingested to iRODS. The extracted descriptive metadata is packaged along with the technical metadata, as a METS document and registered with the iRODS metadata catalog [13]. This process is being implemented in McFarland's collection that for a long time has used a systematic naming convention including image title/terms, its geographical location and type of camera codes, and version control number. To access his files, he may search by any of these elements.

3.3.5 Web Access and Services

Corral provides multiple web servers and supports most popular web application languages, including PHP, Java/Tomcat, and Python. Applications are hosted within a shared environment to minimize administrative overhead, although many collections utilize virtual domains within the web server. Data stored within the iRODS environment can be made openly accessible via a well-defined URL, or can be password protected and accessed via WebDav. Because the same data can be made accessible at high performance from several server nodes, both file-centered web services and web applications are configured for automatic failover across multiple nodes, thus ensuring a high level of

availability. Examples of data collections websites hosted on Corral include OdonataCentral (www.odonatacentral.org) and Fishes of Texas (www.fishesoftexas.org), both of which utilize image and file services, dynamic web applications, and databases to manage catalogs of specimen data.

3.3.6 Curation

Data curation activities happen at the domain science level and at the general collections level. Researchers as curators gather, analyze, interpret, edit, and preserve the data that through those processes becomes a collection, and DMC members technically enable these activities. General curation services include establishing agreements with researchers and tracking and managing services and collections. Researchers determine when the collection is finalized and decide how to provide access, and we work with them to define the appropriate protocols for open access or to implement access restrictions.

4. ADMINISTRATIVE MODEL

As a service and research organization, TACC offers up to five terabytes of free storage space and basic collection services to researchers on campus, and there is a fee structure for collections requiring more storage space. To support complex collection services, the group faces the same limitations and possibilities as the researchers that create the collections. Thus, the group participates from grant proposals with research partners, and provides services in exchange for funding staff hours. In addition, campus organizations use the data facilities as a dark archive for annual fees, which are used to purchase hardware. TACC's cyberinfrastructure intends to surpass the uncertainties of future research funding by embracing the notion that if a collection is built soundly, it will be used and supported, or it can be easily transferred to other archives or managed within other systems. The data storage facility is now entering its 4th year in production and is planned to grow by at least 5 Petabytes of capacity over the next year. There are currently ~500TB of data stored on Corral, 43TB are under iRODS management, and over 40TB of data are in MySQL databases alone. We also have 52TB of data on Ranch, the majority replicas of the data under iRODS management.

5. DISCUSSION AND FUTURE WORK

Acknowledging the evolving nature of data collections, of the research process, and of computing technologies, we present cyberinfrastructure that supports the development and management of sustainable collections. Conducting collection activities within a consistent and flexible data-intensive environment, streamlines the research workflow and enables the implementation of unique functionalities that enhance collections use and reuse. Importantly, it protects the data during those processes and beyond, and eliminates issues of scale for conducting these processes.

To handle the growing number of collections and to better accomplish general curation activities we are developing a collection's catalogue. The database schema is based on Data Documentation Initiative (DDI) and other metadata standards, as well as on elements that we created specifically to trace events (such as those governed by rules and others) and services. The catalogue will also provide an interface to fulfill collections agreements.

The limitations to this cyberinfrastructure are not technical but administrative. Not a library or an archive with the mandate to acquire and preserve collections indefinitely, TACC can provide long-term preservation services for as long as there is an agreement with the collection's curator. For example, the Institute of Classical Archaeology indicated that when and if the Institute cannot support the collection, it should be transferred to the custody of UT General Libraries.

And yet, while libraries and archives are acquiring data collections, they don't have yet the cyberinfrastructure for evolving collections nor the capabilities to host and service very large ones. We are currently collaborating with the UT Libraries and UT System to combine our mutual capabilities in service of the long-term preservation of evolving research collections.

6. REFERENCES

[1] National Science Board. 2005. Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. National Science Foundation. Retrieved 7/15/2011 from, http://www.nsf.gov/pubs/2005/nsb0540/

[2] National Science Foundation. Dissemination and Sharing of Research Results. Retrieved 7/15/2011 from: http://www.nsf.gov/bfa/dias/policy/dmp.jsp

[3] Esteva, M., Trelogan, J. Rabinowitz, A., Walling, D. Pipkin. S. 2010. From the Site to Long-Term Preservation: A Reflexive System to Manage and Archive Digital Archaeological Data. Proceedings of the IS&T's Archiving 2010 Conference, June 1-4, 2010, Den Haag, The Netherlands. Retrieved 7/15/2011 from: http://www.imaging.org/IST/store/epub.cfm?abstrid=43763

[4] McKemmish, S. 1997. "Yesterday, Today and Tomorrow: A Continuum of Responsibility." In Records Continuum Research Group, Retrieved 7/15/2011 from: http://www.sims.monash.edu.au/research/rcrg/publications/recordscontinuum/smckp2.html

[5] DuraCloud. Legacy Documentation. Retrieved 7/15/2011 from:https://wiki.duraspace.org/display/DURACLOUD/DuraCloud+Legacy+Documentation

[6] MacKenzie S. 2010. Managing Research Data at MIT: Growing the Curation Community one Institution at a Time. Keynote IDCC 10, 6-8 December 2010, Chicago, USA. Retrieved 7/15/2011 from: http://www.vimeo.com/17662208

[7] Archaeology Recording Kit. Retrieved 7/15/2011 from: http://ark.lparchaeology.com

[8] Specify Software Project. Retrieved 7/15/2011 from: http://www.specifysoftware.org

[9] iRODS. Data Grids, Digital Libraries, Persistent Archives and Real time Data Systems. Retrieved 7/15/2011 from, https://www.irods.org/index.php/IRODS:Data_Grids,_Digital_Libraries,_Persistent_Archives,_and_Real-time_Data_Systems

[10] Robert McDonald, Chris Jordan, et al..2011. An iRODS Based Data Replication Service for Institutional Data Curation. Poster presentation at the 6th International Conference on Digital Curation, 6 – 8 December in Chicago USA.

[11] Dubrow, A. 2009. Urgent Computing Aids Haiti Relief Effort. Retrieved 07/06/2011, from: http://cms.tacc.utexas.edu/news/feature-stories/2009/urgent-computing-aids-haiti-relief-effort/

[12] FITS, File Information Toolset. Retrieved 7/15/2011 from: http://code.google.com/p/fits/wiki/tools

[13] Walling, D. Esteva, M. 2010. Automating the Extraction of Metadata From Archaeological Data Using iRods Rules, 6th IDCC 10, 6 – 8 December, Chicago USA. To be published in the International Journal of Digital Curation.

A cost model for small scale automated digital preservation archives

Stephan Strodl
Secure Business Austria
Favoritenstrasse 16, 1040 Vienna
Vienna, Austria
sstrodl@sba-research.org

Andreas Rauber
Vienna University of Technology
Favoritenstrasse 9-11/188, 1040 Vienna
Vienna, Austria
rauber@ifs.tuwien.ac.at

ABSTRACT

Assessing the costs of preserving a digital data collection in the long term is a challenging task. The lifecycle costs consist of several cost factors. Some of them are difficult to identify and to break down. In this paper we present a cost model especially for small scale automated digital preservation software system.

The cost model allows institutions with limited expertise in data curation to assess the costs for preserving their digital data in the long run. It provides a simple to use methodology that considers the individual characteristics of different settings. The cost model provided detailed formulas to calculate the expenses. The model supports the detailed calculation of the expenses for the near future and helps to identify the cost trend in the medium and long run (e.g. 5, 10 or 20 years) of the archive. The model monetary assesses the user's work, the purchases of storage hardware and other costs of preserving a digital collection.

In this paper the first version of the model is presented. It includes a discussion about the cost items and presents the calculation the costs. A case study shows the application of the model for a small business setting.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.7 Digital Libraries

General Terms

ECONOMICS, MEASUREMENT

Keywords

Cost model, Digital preservation, Automated archiving

1. INTRODUCTION

Costs are an important aspect in operating a long term archive. Appropriate methodologies and models are required to calculate the cost for medium and long term. The digital information created and managed by institutions is becoming more important for the long term, particularly information that is born-digital and has no analogue counterpart. Examples are business data, construction drawings, patents or data of clinical trials. Digital preservation - ensuring the accessibility and usability of digital information over time - is becoming of broader interests for a wide range of institutions. In the early stages of digital preservation mainly heritage institutions (archives, museum and libraries) were dealing with this issue and had preservation systems in place for their digital collections. Nowadays large organisations and increasing numbers of small institutions are starting or planning preservation activities.

Increased efforts were made in development of small scale and automated preservation archives in the last years. Institutions with limited in-house resources and expertise in digital preservation demand solutions for their digital assets. Solutions are needed that are easy to handle without profound background knowledge. The trend of the developments is toward automation of digital preservation tasks by using knowledge base or recommendation services for decisions.

Digital preservation is a complex continuous process consisting of logical preservation and bit preservation. Current recording media for digital materials are vulnerable to deterioration and catastrophic loss. More challenging than media deterioration is the problem of obsolescence in playback technology. The rapid innovations in computer hardware and software industry result in new storage products and methods on a regular basis. These new products replace the old storage devices and media and hardly ever provide fully backwards compatibility. Beside the physical obsolescence the logical obsolescence of the digital data is often neglected. The rapid development of file formats and the strong dependency between digital objects and the software environment is becoming a pressing problem for archiving. Examples are the periodic release of new office software including new formats for office documents. Other examples are video files that require specific installed encoding software to render the video information. Digital preservation includes all activities to overcome the physical as well as the logical obsolescence. Prominent preservation strategies are migration (to newer storage media (bit preservation) or formats (logical preservation)) and emulation.

iPRES2011, Nov. 1–4, 2011, Singapore.

An early stage issue of all digital preservation systems are the costs. The costs of the next few years are of interest for the management and investors as well as the cost trend in the long term. The total lifecycle costs for preserving a digital data collection consists of several cost factors. Some of them are difficult to identify and to break down. It includes for example user's work of starting a backup process, recurring cost for replacing storage media after their lifespan or cost for migration of the data collection. A challenge particularly for costs calculation for long term preservation is the development of cost factors over time. For example, technological progress reduces the storage costs over time. The data collections on the other hand will grow and also labour costs change over the years. All these developments have to be considered for a potential cost model. Furthermore, the model must consider the characteristics of the different settings including collections and storage media. Storage media for example have different life cycles. Another challenge for a cost model is the quantification of work done by the user. The duration of user tasks varies depending on the skills of the user and the requirements of the setting. A suitable cost model needs flexibility to consider the different characteristics of given settings.

In this paper, a cost model for automated, small scale digital preservation archives is designed. The typical scenarios for this model are small office and home office (SOHOs) settings with a small collection of valuable digital assets for the long term (e.g. business data, construction drawings, models or measured data). In this context, small scale means that the size of the archive collection is small enough to be stored on off-the-shelf storage media (such as external hard discs or DVDs). Larger storage facilities (e.g. tape robot, distributed storage) that required additional management and maintenance effort are not the focus of this model. The model allows calculating the total cost of ownership of preserving a specific data collection over time. It considers the individual characteristics of collections and requirements of the host institution. The here presented cost model is designed for an automated archiving system that automated some archiving tasks, for example the acquisition of data or the backup of the data on storage media. Furthermore, we assume users with limited expertise in digital archiving and preservation. The system needs to obtain the required knowledge from a third party (e.g. knowledge database). In this model we assume a vendor providing the archiving software and the required knowledge as a service. The here presented model considers the cost for the institution that operates the archive.

The Life model [2] was taken as a basis for the cost model. The Life project is a collaboration between University College London (UCL) Library Services and the British Library. It has developed a methodology to calculate the costs of preserving digital information. The methodology provides a very detailed listing of cost items that apply to digital collections throughout their lifecycle. The Life project is focused on professional environments and large institutions. In this paper the cost items of the Life project were analysed how far they apply to an automated preservation system. Where required the model was extended and adjusted for the specific settings.

In this paper we presented a first version of the cost model. It should enable organisation to effectively plan the costs of preserving their digital holdings. The model enables users to calculate the detailed costs of preserving a digital collection for the near future and indicates the cost trend in the long run of an archive. The model assess the activities the users activities carried out, the storage hardware and other costs related for the preservation of a digital collection.

The remainder of this paper is structured as follows. Chapter 2 points out related activities and introduces the Life methodology. Section 3 presents the cost model for automated archiving system. It includes the results from the breakdown of the Life model for automated preservation system. In this section, we further presents the cost model in more detail including the description of the cost elements. A case study in Section 4 presents the cost calculation for a small office setting. Finally, Section 5 draws the conclusions.

2. RELATED WORK

This chapter points out related activities in the field of cost models. Previous efforts in developing cost models for digital preservation are presented. It shows the origins and the motivation behind the preliminary work that resulted in the Life methodology. The Life model forms the basis of the here presented cost model for automated digital preservation archives. A short introduction to current developments of automated preservation systems is also presented in this section.

A first study on costs of digital preservation was done by Tony Hendley in 1998 [10]. The study was sponsored by the British Library and JISC. It provided a first discussion about cost of digital preservation aside storage cost issues that was dominant at that time. A list of data types was defined and a decision model for appropriate preservation methods for the data types was introduced. The proposed cost model defined the cost items of seven modules (creation, selection/evaluation, data management, resource disclosure, data use, data preservation and data use/rights). The cost items are described and discussed in the report but not quantified.

In 1999 Kevin Ashley published an article at the DLM Forum'99 about costs involved in digital preservation [1]. The article stated that the primary influences for the cost are the activities in the archive (such as acquisition, preservation and access) rather than the quantity of the data.

An article about costs focused on logical preservation was published in 2000 by Stewart Granger [9]. He identified three main aspects determining costs of an archive: 'content, data types & formats', 'access' and 'authority & control'. The more these aspects are complex, the more expensive they are. The report provided a first analyse of connection between the costs of digital preservation and the OAIS model [13].

The ERPANET Project published a 'cost orientation tool' for digital preservation [7]. It identified a list of cost factors that should be taken into consideration for digital preservation projects. The factors are arranged around people, digital objects, laws and policies, standards, methods and

practices, technology and systems, and organisation. The factors are discussed in the report but no calculation is provided.

Within the InterPARES 1[1] project a good overview about cost models in digital preservation was published by Shelby Sanett in [17]. Based on a preservation process model of InterPARES a cost model was developed. The costs were organised according to three categories: costs of preserving electronic records, cost for use and user populations. The model strongly focuses on digital records and provided a structure of cost items rather than a calculation model.

Real world studies on costs of digital preservation were conducted by the National Archive of the Netherlands within 'Digitale Bewaring Project' in 2005 [16]. The studies were focused on large archives of government agencies. Based on Testbed studies cost indicators which influence the total costs of preservation were identified. The studies were focused on large archives of government agencies. A first computational model was prepared in form of a spreadsheet.

A study about the costs for preserving research data in UK universities were conducted within the 'Keeping research data safe project'. A series of case studies was executed involving Cambridge University, King's College London, Southampton University, and the Archaeology Data Service at York University [3]. A framework and guidance for determining costs was developed [4]. The model strongly focuses on institutional archiving of research data. The results cannot be directly used in the cost model for automated systems. In the conducted case studies a number of real life data about digital preservation were captured. These data helped to specify the model variables of the here presented cost model (see Section 3.3).

The Life project[2] is a collaboration between University College London (UCL) and the British Library. The aim of the project is the development of a methodology to model and calculation the costs of preserving digital information for the next 5, 10 or 20 years. Within the Life project Watson published a review of existing lifecycle models and digital preservation [21]. The review is focused on library sector and forms the basis for the Life methodology. The Life project consists of three phases. The first phase (Life v1) of the project ran from 2005 to 2006. Based on the review [21] a first version of the Life model was developed [15]. The model breaks the costs down into six main lifecycle categories. In the second phase of the project the model was validated by an economic review [5]. Based on feedback received on Life v1 and the economic review an updated version of the Life cost model (Life model v2) was published [2]. The elements were described in more detail and sub-elements were suggested. The Life model v2 was taken as a basis for the here presented cost model (as described in Section 3.3). The recommendations from the economic review were considered in this work for example the handling of inflation for different goods (e.g. wages, media). The generic model of the Life methodology was used as guidance for the formula of the cost model provided in Section 3. In 2009 the third phase of the Life project started. The aim is the development of a predictive costing tool [11]. The Life model was most suitable basis for a cost model of automated archiving systems.

A number of research initiatives have emerged in the last decade in the field of digital preservation, mainly carried out by memory institutions. Automation of preservation processes has been identified as one of the great challenges within the field of digital preservation (e.g. in the DPE roadmap [6]). A few projects have already addressed the automation of components of a preservation archive.

The CRIB project [8] for example has developed a Service Oriented Architecture implementing automated migration support. The digital objects are transferred to a server infrastructure and migrated objects are returned. The actual migrations of the objects are executed on the server side. CRIB is integrated into the RODA repository[3].

The Panic Project [12] developed a framework to dynamically discover suitable preservation strategies. Panic uses semantic web technologies to make preservation software modules available as Web services. The system is designed for large-scale repositories that implement the required services invoker.

The PreScan system [14] automatically extracts embedded metadata from digital objects. The system scans objects on a hard disc and manages their metadata in an external repository that supports Semantic Web technologies. The metadata could be used to implement digital preservation support.

The Hoppla archive [18] provides a (semi-) automated preservation archive for small institutions. The system combines back-up and fully automated migration services. It provides a high degree of automation for a wide set of functions of the archive. The components of Hoppla include automated acquisition, ingest, data managers, preservation management, access and storage. The concept and the design of Hoppla are presented in more detail in [18].

3. COST MODEL FOR AUTOMATED PRESERVATION ARCHIVES

In this section the cost model for automated digital preservation system is presented. The model was designed on the basis of the Life model v2 [2]. Some assumptions and conditions are required for the model that are described in Section 3.1. Based on these assumptions the Life model was analysed to which extent it is applicable for a small scale automated preservation system. The result of the analysis is presented in Section 3.2. As the Life model does not fully support the specific setting of automated preservation system the model is extended and adjusted where required. The resulting cost model is presented in Section 3.3 in detail.

3.1 Assumption and conditions

Some assumptions and conditions regarding to the environment and the archiving system have to be defined for the

[1] http://www.interpares.org
[2] http://www.life.ac.uk
[3] http://roda.di.uminho.pt

cost model. Settings where these assumptions and conditions are not fulfilled have to be considered separately.

- **Small scale data collection**

  The first condition concerns the collection size. The cost model focuses on small scale data collections that can be stored on off-the-shelf storage media (e.g. external hard discs or DVDs). Settings with data volumes that require special maintained and customised storage infrastructure (such as storage server, tape robots, etc.) are not covered within the parameters provided for this model.

- **Licensing & Rights of the data**

  The rights management is not within the scope of this cost model. We proceed on the assumption that the institution owns the content and holds all required rights and licenses to process, manipulate and store the data.

- **(Semi-)Automation preservation system**

  The here presented cost model is designed for an archiving system that executes archiving tasks automatically, for example the acquisition from data carriers, characterisation, migrations and storage. An example of an automated preservation system is the Hoppla system [18].

- **Outsourcing of knowledge and expertise in digital preservation**

  We assume that the archiving system is operated by an institution that has no profound knowledge of digital preservation and not the resources available to acquire it in-house.

  The system needs to obtain the required knowledge and expertise from somewhere else, e.g. a knowledge database, or a web service operated by experts. Moreover the system has to automatically take decisions and give recommendations to the user. The cost of the creation, operations and maintenance of the knowledge services needs to be considered in the cost model (e.g. in the form of a licence fee).

- **No dedicated archiving host system**
  The here considered automated archiving systems have typically only very basic hardware requirements for host systems. We assume that in small institutions the archiving system usually shares the hardware with other operative systems (storage server, etc.) and no dedicated hardware is needed. Thus we do not consider the hardware of the host system in the cost model, except from, obviously, the actual storage media.

- **Internal archive**

  The preserved content is for internal use and billing and access to external customers is not within the scope of the model.

3.2 Life Cost Items applied to automated Archiving Systems

The cost items of the Life model were analysed to which extent they are applicable for a small scale automated preservation system. As the Life model is designed on a generic level not all of the cost item are relevant for an automated system.

Moreover not all cost items that are applicable to an automated system actually incur direct costs. The system automates lots of activities listed in the Life model (e.g. obtaining of data or access provision). We use the Life model v2 in this work. Based on conditions defined in Section 3.1 the Life model was analysed. Due to the limited available space in this paper we can only present the results of this work, a more in-depth discussion about the applicability of the cost items can be found in 'Cost model for automated archiving system' [4].

The result of the evaluation is shown in Figure 1. For all cost items of the Life methodology we determine whether they are

- not applicable/relevant for an automated system [NR] or
- no direct costs incur as the activity is executed by the archive system software [NC] or
- user work or purchasing is needed. The cost item has to be considered in the cost model [CM]. We further distinguish between the client side [CM/C] and the server side [CM/S].

Some cost items in Figure 1 have two entries. In this case sub-elements have different assignments. In this work we only interested in the client side of the archive system. For the server side, we assume an update service for the archiving software system that provides the required knowledge and services. The costs for these activities are indirect paid by the client via the software system and an annual fee.

Other cost models were also analysed how far the support automated archiving system and whether all expenses are covered by the Life methodology. As a result of this work, the cost model was extended by the costs for the archiving software. The resulting model is presented in the next section.

3.3 Cost model

A cost model for a small scale automated preservation system must be flexible enough and open enough to consider the individual characteristics of the different settings. The characteristics include amongst others the collection, the used storage media, the requirements and the effort spend by the user for tasks. Otherwise the model should be as specific as possible to serve users with limited expertise as a guide to calculate the costs for preserving their digital holdings.

[4] http://www.ifs.tuwien.ac.at/~strodl/paper/techreport_costmodel.pdf

Acquisition	Ingest	Bit-stream Preservation	Content Preservation	Access
Selection [CM/C] [NC]	Quality Assurance [CM/C]	Repository Administration [CM/C] [NC]	Preservation Watch [CM/S] [NC]	Access Provision [NC]
Submission Agreement [NR]	Metadata [CM/C]	Storage Provision [CM/C]	Preservation Planning [CM/S] [NC]	Access Control [NR]
IPR & Licensing [NR]	Deposit [NC]	Refreshment [CM/C]	Preservation Action [CM/C] [NC]	User Support [CM/S]
Ordering & Invoicing [NR]	Holdings Update [CM/C]	Backup [CM/C] [NC]	Re-ingest [NC]	
Obtaining [NR]	Reference Linking [NC]	Inspection [NC] [NR]	Disposal [CM/C]	
Check-in [NC]				

Figure 1: Life Model applied to automated archiving system

Based on the analysis of the Life model and other cost models, the here presented cost model was designed. The structure and the cost items of the model are shown in Figure 2. The Life model was extended by the category 'preservation system'. It contains the costs for the archiving preservation software system including update service and potential customisations of the software. The structure of bit-stream preservation cost items is more detailed as in the original Life model. Few cost items in this model are optional. Their use depends on the actual setting and the used software system. Optional cost items are marked with an asterisk in Figure 2.

The cost model provides formulas to calculate the costs of the single cost item. One of the basic principles of the model is the modular structure. The cost of a single item can be calculated separately. The suggested formulas can be easily adjusted or replaced by actual costs or other models for cost calculation. The suggested formulas should provide a starting point to assess the cost for an archiving system. The cost model deals with three types of costs: manual work that has to be done by the user, purchases of hardware storage (such as storage media) and other expenses (e.g. software fees, online storage services). The monetary assessment of these factors allows the calculation of costs for preserving a digital collection for the institutions.

An aim of the model is the assessment of user work. It is a challenging task as the work strongly depends on the user, the collection, the archival system and the requirements. In order to assess the user work the model considers different level of preservation requirements for a given setting. Depending on requirements the user will put more or less effort in preserving the collection and therefore investing more time in executing preservation tasks. The model further introduces a calculation to estimate the errors that occur during migration and backup process. Based on the error rate the effort for monitoring the process and fixing problems can be estimated.

The cost model provides calculation for the hardware storage demand of the archive. It considers the growth of the collection, hardware migration (replacement of old media after their life span) and the cost trend of storage media. In the model different storage media types are supported including online storage.

In order to support different settings, the model comprises optional effort and cost items. Example for optional effort is metadata assignment by the user. It incurs expenses, but it is not mandatory and optional for the user. Another example for optional costs is customisation of the archival system. In order to fulfil legal obligations or strict requirements the adoption and customisation of the software system can be required. The model takes these expenses into account.

As the cost model deals with expenses in the distant future we need to consider the cost trends over time. In order to calculate the costs of future investments the time value of money needs to be considered. In our model we use real prices that are inflation-adjusted prices, where prices of different years are divided by the general price index for the same year. It allows the comparison of prices over the years and the identification of cost trends. For a long term archive two important costs factors change significantly over time with another long-term trend than general price index, first the costs of storage and the cost of labour work. Both developments are considered in the cost model.

The model supports can be used for existing archives as well as for planned ones. Year 0 (t=0) is the first year of the archive in the model, it is used for archives built from scratch. In this year the initial setup of the archive is done. Additional effort for the set up is considered, especially for user settings such as policies and data selection. In cost calculation for already existing archives year 0 is skipped and the calculation starts with year 1.

The model provides detailed formulas for the cost items. Due to the limited available space in this paper we present the basic concepts of the formulas and the calculation. The detailed formulas are shown in Figure 2. They are in brackets within the text. Some of the variables used in the model will be explained in the following description. A detailed discussion of all cost factors in presented in 'Cost model for automated archiving system' [5].

The cost calculation for long term archives depends on many input factors. There are two kinds of variables used in the cost model, model variables representing common measurements and cost factors that are individual for each setting. The model variables strongly depend on the used archive software. They include the expected duration for users activities such selection of data sources, storage procurement, setting policies, etc. Model variables are predefined and are quite similar for most of preservation settings. The second type of variables in the cost model is cost factors that are individual for each setting and need to be defined from the user. They include for example size of the collection, the expected growth rate and the costs of manual work.

There are few key figures that are used in a number of formulas that describe the setting. The size of the collection

[5] `http://www.ifs.tuwien.ac.at/~strodl/paper/techreport_costmodel.pdf`

stored in the archive (sc(t)) is calculated for every year based on a starting size and a yearly growth rate. The collection growth includes new added objects, migrated objects and stored history of changed objects.

The number of objects in the collection (noc(t)) is used for the error calculation for backup and migration. The number is also calculated for each year based on a starting number and a growth rate per year. In order to monetary assess the users work a cost for manual work per hour (cwh(t)) need to be set by the user. A yearly salary adjustment rate is used to consider the cost trend of salaries over time.

Another important factor used in this model is the user requirement level (nur). Depending on the setting and the relevance of the data collection the user will put more or less effort in preserving the collection and therefore investing more time in executing preservation tasks. The user requirement level specifies a scale that represents a multiplication factor for the effort.

In the following section the cost items of the cost model as shown in Figure 2 are presented.

3.3.1 Client total cost (cto)

The overall costs of preserving a digital collection (cto(t)) are the sum of all cost items. All cost items and the formulas are shown in Figure 2.

$$cto(t) = csp(t) + cse(t) + cmc(t) + chu(t) + csh(t) + cre(t) + csp(t) + cdr_t + csu_t + cbp(t) + cba(t) + cqp(t) + cdi_t + css_t + ccs_t$$

3.3.2 Acquisition

The acquisition includes the selection of the policies (csp(t)) and the selection of the content (cse(t)). Both activities have an initial effort in the first year of an archive. Automated archiving systems usually provide predefined policy profiles. It should help the users to select an appropriate policy for their needs. In the first year the data sources for the collection need to be initially selected (including selecting the sources and the settings of the filter criteria). In both cases we expect that settings with more detailed requirements will spend more effort adjusting the policies and selecting the content. The effort is multiplied by the user requirements level (nur). The effort for selecting the policies and content strongly depends on the archive software. They are defined as model variables. A review of both settings is planned on yearly basis.

3.3.3 Ingest

The ingest includes the optional cost item 'metadata creation' (cmc(t)) and 'update holing' (chu(t)). Automated preservation systems automatically collect and assign metadata to the objects in the repository. In many cases the manual assignment of metadata can improve the usage of the collection (e.g. statistics and search). Due to the labour-intensive work, the metadata assignment can cause considerable costs. The costs are calculated by the optional metadata creation effort per year (defined as emc_t) multiplied by the hourly rate of the user.

The update of the holdings is performed by the archive software. User effort is required to start the update process and prepare the setting. The user needs to start the application and make all sources and storage media available. The effort strongly depends on the software and the expected effort is defined in a model variable. The effort is multiplied with the number of ingests per year (nic).

3.3.4 Bit-stream Preservation

Bit-stream preservation is a core cost component of long term preservation. It covers the cost of the hardware and the manual work for physical backups (see Figure 2).

In the model we distinguish between three types of bit-stream media: re-write media (such as HD) (abbr. rw), write once media (such as CD, DVD) (abbr. wo) and online (e.g. SSH, web services) (abbr. on). In the model we use bm ... for all bit-stream media, bmh ... for all hardware media (re-write and write once media), and bmo ... for online media. The model can be easily adjusted and enhanced by adding new media. The cost model further supports multiple separate copies of the data collection per storage media (for example two online storage services, or three separate copies on hard discs). The number of separate copies is defined as Backup Level for each media (bl_{bm}).

Storage hardware (csh)

The storage hardware represents the main cost item of bit-stream perseveration. We distinguish for the storage hardware between storage as a service (e.g. online storage) and storage on hardware (e.g. re-write media, write once media).

New innovation and continuous development of storage technology steadily increases the storage capacities and decreases the cost for storage. In order to consider the development of storage media we introduce a storage cost deflator rate. The rate is defined for each media and defines the annual improvement of the storage capacity per year in percentage (rmd_{bm}). The storage prices are calculated for every year. The development of the storage prices is not constant every year depending on technological progress and innovation. But we have a look in the past, a constant curve of price decreases provides a good approximation (with few outliers) of the storage development in the long run [20].

For storage as a service we have yearly expenses. The collection size is multiplied by the current storage costs for the service. The result is multiplied by the number of separate online storages (backup level).

The expenses for storage on hardware cover the refreshment of storage media (re-write and write once media (bmh)). In order to avoid physical data loss the storage media have to be refreshed after their expected life time. The variable refreshment cycle of a media (rc_{bmh}) defines the expected life time of the media. Due to the different refreshment cycles the storage hardware costs vary every year and have to be calculated for each year individually. The function frc(t,rc_{bmh}) defines the years of storage migration. In order to calculate the costs for a replacement of a storage media the required size of the new storage media has to be calculated. As the collection size grows over time the storage medium need to have enough capacity to store the collection up to next refreshment cycle. The size is multiplied by current storage prices and by the number of separate copies.

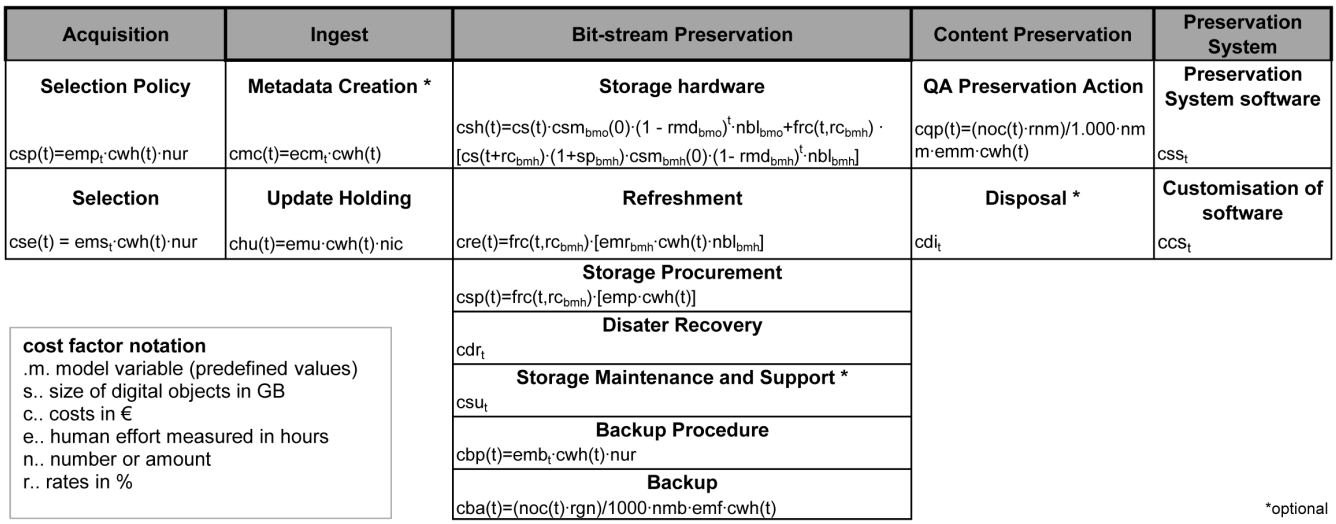


Figure 2: Cost model for small scale automated digital preservation archives including formulas for the cost items

As initial set-up all storage hardware is bought at year 0 of the archive. After that the media are replaced according their refreshment cycles.

Refreshment (cre)
The replacement of old storage media (storage migration) requires in addition to new storage hardware also manual work. The migration process is executed by the software, but the user needs to set up the environment and start the migration process. The migration is a very critical task as the complete collection is transferred to a new medium. The correctness of the migration is essential to ensure the availability of the data. The checking, analysing the report and error logs of the migration is critical and requires most of the time. The effort depends on the software and the number of separate copies.

Storage procurement (csp)
Additional to the hardware and refreshment costs the procurement of the new storage hardware causes expenses. Only minimal effort is estimated as the internet suppliers ease the procurement procedure for the user.

Disaster Recovery (cdr)
Backup copies stored on same location do not help in case of natural disasters such as fire or flood. It is strongly recommended to keep a copy of the data on an off-site location. The cost model deals with the disaster recovery for the data, the recovery of the infrastructure is out of the scope of this model. An example for an off-site location is a safe deposit box. The costs for disaster recovery are individual for each setting depending on the strategy and have to be specified by the user. The use of online storage could also be a practicable disaster recovery strategy. In this case the costs are coved as storage hardware (storage as a service).

Storage Maintenance and Support (optional) (csm)
Institutions that operate a small scale digital preservation archive do not tend to have maintenance and support contracts for their storage devices. It is an optional cost item in the cost model.

Backup Procedure (cbp)
The backup procedure is guided by backup policy. In year 0 of the archive the initial backup policy needs to be defined by the user. Automated archiving system helps user with predefined profiles for the policy selection. Thus a minimal effort is assumed for this activity. User with higher requirements will invest more time in defining their backup policy in more detail.

Backup/ Backup monitoring (cba)
The backup action is executed by the archive software. Automated backups tend to be error-prone tasks. The user needs to analyse the logs and reports of the process. If necessary the user needs to fix problems (e.g. restart process, re-insert external devices, etc.).

We calculate the expected effort for log analysis and error fixing on the assumption that the probability of errors during backup correlates with the number of new objects backup-ed in the collection. The larger the collection the more errors occur. A mean failure backup rate is defined per 1.000 objects(nmb). They error rate will depend on the setting (the used hardware, software and the users). Expertise from similar setting can be provided guidance values for the error rate. Based on the number of new objects added to the archive per year, the error rate and estimated time to fix the failures the effort for Backup /Backup monitoring is calculated.

3.3.5 Content Preservation

Quality assurance of the preservation actions in the archive is a key aspect of all digital preservation system. As migration (preferred content preservation action for automated archives) is a modification of the data the validation of the results is important to guarantee the trustworthiness of the

archive.

The automation of migration validation is key challenge of digital preservation. Part of the work has to be done by the user (e.g. analysing logs). Similar to the backup cost, a mean failure rate is used to calculate the user effort. The mean migration failure rate is defined as a number of failed migrations per 1.000 executed migrations (nmm). The failure rate depends on complexity of formats and accuracy of the used migration tools. Work on the complexity of file formats was done in the Generic Life Preservation model (Section 8.4.8 in [15]). The File Format Complexity scale can be used to adjust failure rate. The number of migrations executed in the archive depends on the number of elements in the archive (noc(t)) and a migration rate (rnm). The migration rate depends on formats in the collection and the user settings.

The time spend by user for QA preservation actions (eqa(t)) is calculated by the mean failure migration rate (nfm(t)) multiplied by number of migrations per year the estimated time to analyse and fix the failure. The result is multiplied by the hourly rate of the user.

An optional cost item of 'content preservation' is disposal. The disposal of digital objects from a collection strongly depends on the collection and the used storage media. The expenses for disposal need to be specified for each setting (cdi_t).

3.3.6 *Preservation System Software*

In this work the ordinal life model was extended by the costs of the preservation software system. The preservation system software includes two cost items, the costs of the digital preservation software system and customisation of the software.

The initial costs for the archive software are booked in year 0 of the archive. We expect annual costs for update and maintenance service (e.g. new preservation rules). The required update service strongly depends on the host institution, its requirements and obligations, the collection and the expertise in-house.

Individual requirements and obligation of institution can require customisation and adoption of the archive software (for example support of specific formats, integration of specific tools).

The costs for the customisation for each year are captured in this cost item 'Customisation of system' (ccs_t). The customisation is specific for each setting and can vary from year to year. Settings with higher preservation requirements tend to have higher spending for the customisation than settings with basic preservation requirements. This cost item has to be set by the user.

We identified four potential areas for customisation of a digital preservation system with respect to technical functionality: quality assurance of objects, metadata creation, integration of new preservation solution and quality assurance of preservation action. Other customisation can include for example the integration of the archive into existing systems or connection to specific data sources or storage systems. The adoption of the user interface is also a typical customisation request.

4. CASE STUDY

A first case study shows the cost calculation by using the proposed model for a small business setting. The business wants to preserve selected data of the business activities over time. There are no legal obligations for preserving, but the data are needed for later analysis and reuse. The data consists in the main of common office documents and images. The archive is built from scratch and a first cost estimation should be done for the short term. Moreover, the cost trend of a potential archive in the long run should be calculated. The model variables used in this case study are based on our experience with the Hoppla archiving system [18].

The initial collection has a size of 75GB. We expect a rather slow growth of 5% every year of the collection. Two ingests are planned every year. The archive data are stored on two separate external hard discs. They are replaced every five years. One backup copy is made on optical write once media. In order to off-side location copy of the archive data an online storage service is used.

The user has only basic requirements and only formats that are at immediate risk of becoming obsolete are migrated. The hourly rate of the user is €70. An increase of 1,5% every year is assumed for the hourly rate. As we use inflation-adjusted prices in the model, the increase of the hourly rate is additionally to the inflation. For the preservation software an off-the-self preservation system is planned with initial costs of €140 and annual service fee of €30 for updates of the preservation rules.

In this case study additional metadata are assigned to the collection by the user. The data are manually categorised to enhance search functionalities and statistics. After each ingest the user assignees categories to the new data. A few hours are planned for each ingest, for the cost calculation we expect about 13 hours per year for metadata assigning.

Table 1 shows the costs of the single cost items. The total costs per year ranges from about €1500 up to €3500. A visualisation of the total costs is shown in Figure 3. It shows a constant increasing cost trend and some outliers with higher costs than the constant trend.

There are higher expenses in year 0 of the archive. The initial purchase of the hardware and the initial set up of the system (e.g. policies, section of data) cause the additional costs. The outliers in the following years are caused by the replacement of storage media. Every five years the re-write media are replaced by new ones (every four for write once media). The media migration causes additional costs for the new hardware and the effort by the user for the migration. Table 1 shows that the cost item 'refreshment (cre(t))' causes the increase of costs in these years. The cost for the labour work (refreshment) of the hardware migration is much higher than the actual hardware costs.

The constant increase of the cost level is caused by the increase of the hourly rate of the user over the years. This

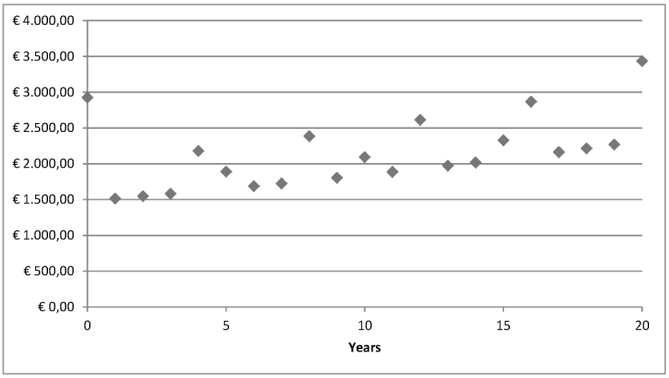


Figure 3: Case study: graph of total cost

trend can be seen in the cost items of acquisition and ingest that consist of constant manual work. The amount of work and costs for hardware keep constant over time. The decrease of the storage costs make up the growth of the collection.

When we have a look into more detail we can see than the highest cost are caused by the manual work of the user. Between 85% and 90% of the costs of the archive are caused labor costs. Table 1 shows that other expenses (such as storage hardware) are causes only a small proportion of the total costs. The biggest single cost factor in the model is the metadata assignment. The estimated workload of the user for preservation activities is between 20 and 30 hours per year. Table 1 further shows that 'Backup/ Backup monitoring (cba(t))' and 'QA Preservation Action (cqp(t))' causes only small expenses in the beginning. With growing collection the costs (and the effort) for quality assurance of the preservation actions strongly increases.

The case study shows that the required software and hardware only reflect a small part of the actual costs of an preservation archive. The manual work of operating the archive causes the majority of the cost.

The model helps to easy identify the major cost items of preserving a collection and provides a good estimation of required resources (work and funds). The perspective of the calculation allows to identify the cost trend of a growing collection.

5. CONCLUSION

The here presented cost model provides a comprehensive methodology to assess the expenses for preserving a digital collection. It aims to provide a simple to use methodology to calculate the cost of a small scale automated digital preservation archives. The cost model comprises all cost items that are relevant for small business preservation setting. The model supports the calculation of the expenses for the near future and also to indicate the cost trend in the long run (e.g. 5, 10 or 20 years). The model is based on the Life model v2. It is adapted and extended for the specific needs of institutions using preservation system that provide a high degree on automation of preservation tasks.

The model provides a modular structure with optional cost items. It can be easily adjusted extended or reduced according actual conditions. Moreover the formulas provided in the model are considering the requirement, obligations and optional effort of different settings.

The cost model is subject to some assumptions and conditions regarding the environment and the archiving system. They help to substantiate the abstract level of common cost models. The model provides detailed cost formulas with measurable input factors.

The model considers three different tree types of costs: work the user has to execute, purchases (such as storage hardware) and other expenses (such as service fees). The model supports the estimation of the user's effort that is required for executing preservation tasks (e.g. selection of content, analysing the report and error logs). Moreover, a model for estimate error rates during migration and backup process is introduced in the cost model. It helps institution to gain a better understanding of the effort and the associated costs of operating a digital archive.

Other expenses of preserving a collection are storage media. The model provides a detailed calculation of the required storage devices. It supports different storage media. Moreover, it considers the lifespan of the media and storage media migrations.

The cost model for small scale automated preservation system provides formulas that assess the user work and expenses of the cost items. This allows to identify expensive and work intensive cost items in preserving a digital collection. The cost model and especially the formulas should provide a starting point for initial assessment of the costs for preserving their digital holdings.

A first case study is presented in this paper. It presents the cost calculation of a small scale office setting for a planned preservation archive. The case study showed the detailed costs calculation for the near future. It allows to identify the major cost factors of running an archive and to estimate the required workload. In this case study about 20 and 30 hours of work are calculated per year. Moreover, the long term cost trend of the planned archive was shown. In the case study the costs keep constant over time with a slightly increase caused by wage increase. The slow growth of the collection has no big impact on the cost development of the archive. The case study shows that the biggest cost factors are the work done by the user. The cost model should help to planned and budget a preservation archive.

More case studies in different settings are necessary to further verify the proposed model. The effects of different software products and storage strategies need to be evaluated in more detail. Another important point for further studies is the relation between effort by the operator and size of the collection. Sufficient real data are need for fine-tuning the model variables. It would further allow the identification of critical factors that affect the time to execute tasks and help improving preservation software system.

With the cost model for small scale automated digital preser-

	Acquistion		Ingest		Bit Stream Preservation							Content Preservation		Preservation		
Year	Select Policy	Selection	Metadata Creation*	Holding Update	Storage hardware	Refreshment	Storage Procurement	Disaster Recovery	Storage Maint. and Support *	Backup Procedure	Backup	QA Pres. Action	Disposal	System software	Customisation	Total SUM
t	csp(t)	cse(t)	cmc(t)	chu(t)	csh(t)	cre(t)	csp(t)	cdr_t	csu_t	cbp(t)	cba(t)	cqp(t)	cdi_t	css_t	ccs_t	
0	140,00	420,00	910,00	210,00	282,46	560,00	70,00	0,00	0,00	35,00	50,40	98,28	0,00	150,00	0,00	2.926,14
1	14,28	35,70	928,20	214,20	172,15	0,00	0,00	0,00	0,00	14,28	2,06	104,26	0,00	30,00	0,00	1.515,13
2	14,57	36,41	946,76	218,48	174,80	0,00	0,00	0,00	0,00	14,57	2,18	110,59	0,00	30,00	0,00	1.548,37
3	14,86	37,14	965,70	222,85	177,34	0,00	0,00	0,00	0,00	14,86	2,31	117,32	0,00	30,00	0,00	1.582,38
4	15,15	37,89	985,01	227,31	251,16	454,62	37,89	0,00	0,00	15,15	2,45	124,45	0,00	30,00	0,00	2.181,09
5	15,46	38,64	1.004,71	231,86	226,79	154,57	38,64	0,00	0,00	15,46	2,60	132,02	0,00	30,00	0,00	1.890,75
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
10	17,07	42,66	1.109,28	255,99	228,07	170,66	42,66	0,00	0,00	17,07	3,50	177,34	0,00	30,00	0,00	2.094,30
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
15	18,84	47,11	1.224,74	282,63	228,92	188,42	47,11	0,00	0,00	18,84	4,70	238,21	0,00	30,00	0,00	2.329,52
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
20	20,80	52,01	1.352,21	312,05	384,56	832,13	104,02	0,00	0,00	20,80	6,31	319,99	0,00	30,00	0,00	3.434,88

Table 1: Case study: Costs calculation of a small business setting (Costs in Euro)

vation archives the cost for preserving a digital collection can be planned in an efficient way. The model has a very modular structure and it is easy to adopt for individual needs. The comparison of the cost for years help to identify cost trends and allows a solid budget and resource planning for a digital preserving archive.

Acknowledgements

Part of this work was co-funded by COMET K1, FFG - Austrian Research Promotion Agency.

6. REFERENCES

[1] Ashley, K. Digital archive costs: Facts and fallacies. In *DLM Forum '99* (Brussels, Belgium, October 1999), E. Commission, Ed.

[2] Ayris, P., Davies, R., McLeod, R., Miao, R., Shenton, H., and Wheatley, P. The LIFE2 Final Project Report. Report, UCL Departments and Research Centres, 2008.

[3] Beagrie, N., Chruszcz, J., and Lavoie, B. Keeping research data safe - a cost model and guidance for uk universities. Tech. rep., JISC, 2008.

[4] Beagrie, N., Lavoie, B., and Woollard, M. Keeping research data safe 2. Tech. rep., JISC, 2010.

[5] Björk, B.-C. Economic evaluation of life methodology. `http://eprints.ucl.ac.uk/7684/`, July 2007.

[6] DigitalPreservationEurope (DPE). Research roadmap. `http://www.digitalpreservationeurope.eu/publications/reports/dpe_research_roadmap_D72.pdf`, October 2007.

[7] erpanet. Cost orientation tool. erpaguidance, ERPANET, 2003.

[8] Ferreira, M., Baptista, A. A., and Ramalho, J. C. An intelligent decision support system for digital preservation. *Int. Journal on Digital Libraries 6*, 4 (July 2007), 295–304.

[9] Granger, S., Russell, K., and Weinberger, E. Cost elements of digital preservation. `http://www.webarchive.org.uk/wayback/archive/20050111000000/http://www.leeds.ac.uk/cedars/colman/costElementsOfDP.doc`, October 2000.

[10] Hendley, T. Comparison of methods & costs of digital preservation. British Library Research and Innovation Report 106, "British Library Research and Innovation Centre", 1998.

[11] Hole, B., Lin, L., McCann, P., and Wheatley, P. Life3: A predictive costing tool for digital collections. In *Proc. of the 7th Int. Conf. on Preservation of Digital Objects (iPRES2010)* (2010), pp. 359–363.

[12] Hunter, J., and Choudhury, S. PANIC - an integrated approach to the preservation of complex digital objects using semantic web services. In *International Journal on Digital Libraries: Special Issue on Complex DigitalObjects. 6 (2).* (Berlin, April 2006), Springer-Verlag, pp. 174–183.

[13] ISO. *Space data and information transfer systems – Open archival information system – Reference model (ISO 14721:2003)*, 2003.

[14] Marketakis, Y., Tzanakis, M., and Tzitzikas, Y. Prescan: towards automating the preservation of digital objects. In *MEDES '09: Proc. of the Int. Conf. on Management of Emergent Digital EcoSystems* (New York, NY, USA, 2009), ACM, pp. 404–411.

[15] McLeod, R., Wheatley, P., and Ayris, P. Lifecycle information for e-literature: full report from the life project. Report, LIFE Project, 2006.

[16] Nationaal Archief. Costs of digital preservation. `http://www.digitaleduurzaamheid.nl/bibliotheek/docs/CoDPv1.pdf`, May 2005.

[17] Sanett, S. Toward developing a framework of cost elements for preserving authentic electronic records into perpetuity. *College & Research Libraries 63*, 5 (September 2002), 388–404.

[18] Strodl, S., Motlik, F., Stadler, K., and Rauber, A. Personal & SOHO archiving. In *Proc. of the 8th ACM IEEE Joint Jonference on Digital Libraries (JCDL'08)* (Pittsburgh PA, USA, 2008), ACM, pp. 115–123.

[19] Strodl, S., Petrov, P., Greifeneder, M., and Rauber, A. Automating logical preservation for small institutions with hoppla. In *Proc. of the 14th European Conf. on Research and Advanced Technology for Digital Libraries (ECDL2010)* (2010), Springer Berlin / Heidelberg, pp. 124–135.

[20] Walter, C. Kryder's law. *Scientific American* (August 2005).

[21] Watson, J. The life project research review: mapping the landscape, riding a life cycle. Tech. rep., LIFE project, November 2005.

Cost Aspects of Ingest and Normalization

Ulla Bøgvad Kejser
The Royal Library
Postbox 2149
1016 Copenhagen K
+45 33 47 47 47
ubk@kb.dk

Anders Bo Nielsen
Danish National Archives
Rigsdagsgården 9
1218 Copenhagen K
+45 33 92 83 26
abn@ra.sa.dk

Alex Thirifays
Danish National Archives
Rigsdagsgården 9
1218 Copenhagen K
+45 33 92 23 69
alt@ra.sa.dk

ABSTRACT
The Danish National Archives, and The Royal Library and the State and University Library are in the process of developing a cost model for digital preservation: Each of the functional entities of the OAIS Reference Model are broken down into measurable, cost-critical activities, and formula are being tailored for each of these in order to create a generic tool for estimating the short and long-term costs of digital preservation. This paper presents an introduction to the subject of the costs of digital preservation and describes the method used to develop the Danish Cost Model for Digital Preservation (CMDP). It then describes how the OAIS functional entity, Ingest, has been included in the model. For institutions basing their digital preservation strategy on migration, a major cost pertaining to Ingest is *normalization*, a digital migration from production to preservation format and structure, which is often quite complex in comparison to the subsequent migrations within the archive. The paper accounts for three aspects of migrations, which are decisive for the costs: the required migration quality, when in the lifecycle the first migration takes place, and how often subsequent migrations are executed. Lastly – with view to increasing the model's precision – existing cost data from submission projects have been used to test the CMDP and the results of this test are described.

Categories and Subject Descriptors
H.3 m [**Information Storage and Retrieval**]: Miscellaneous.

General Terms
Measurement, Documentation, Economics, Standardization.

Keywords
Activity based costing, Cost model, Ingest, Migration, Normalization, OAIS Reference Model, and Preservation.

1. INTRODUCTION
The digital preservation field lacks economic models, i.e. models which account for costs and benefits, to enable justification of investments [2]. In recent years several projects have worked to define cost benefit models, such as the KRDS project [1] and the DANS cost model for digital archiving [7]. Also, some projects have developed cost models *per se*, such as the cost model created by the National Archives of the Netherlands [15], the LIFE Costing Model [9], and the NASA-CET [11].

In Denmark the Ministry of Culture has funded a project to set up a model for costing preservation of digital materials held by national cultural heritage institutions. The project has been undertaken by The Royal Library, The State and University Library, and The Danish National Archives, and consumed a total of 2 Full-time equivalents (FTE). In the first phase of the project (2009) a methodology for a cost model was developed, and designated the Cost Model for Digital Preservation (CMDP) [12]. The CMDP is based on the Open Archival Information System (OAIS) Reference Model [3] and on activity-based costing [5]. Furthermore, the work draws upon general costing principles defined in the International Cost Model Manual [13]. The CMDP is designed to be generic in order to enable calculation, estimation and comparison of the costs of digital preservation across memory and research institutions holding different types of digital materials. So far the model only addresses costs of digital preservation by the migration strategy. With time we envision to enable costing of the emulation strategy.

For developing the CMDP, we used the OAIS model to identify functions and divide them into delineated activities. We then identified the cost parameters (variables) related to the individual activities and operationalised them as formulas in a spreadsheet. Thus the CMDP spreadsheet tool represents modules based on the functional entities in the OAIS Model. In the CMDP costs are stated as the time it takes to perform an activity multiplied by the wage, plus the costs of any provisions. The CMDP only accounts for so-called cost critical activities, defined as activities that take a minimum of one person week to complete. A person week is set to 32 effective working hours, but as other variables in the spreadsheet, such as wages, it may be changed by users depending on local requirements. The CMDP includes all direct expenses of establishing and operating the preservation system as well as indirect costs, such as general administration (overhead). Eventually the model will also take financial adjustments, e.g. inflation, into account. While this is the ideal goal, the task is hard, and it may well be necessary to scale down the ambitions.

The spreadsheet and other documentation are available from the project web site[1].

In the first phase of the project we operationalised costs of the functions under the functional entity Preservation Planning and focused on the costs of the migration strategy. We also operationalised functions from related OAIS functional entities, which sustain Preservation Planning, especially functions under the functional entity Administration.

In the second phase of the project we have addressed the costs of the activities within the OAIS functional entity Ingest and related functions from Administration. To improve the identification of Ingest activities we also analyzed the Producer-Archive Interface Methodology Abstract Standard (PAIMAS) [4], which provides a detailed description of the interactions that take place between the OAIS roles, Producer and Archive. Finally, to account for the costs of normalizations, we have improved the formula for digital migrations developed in the first phase of the project.

We have used OAIS terms as far as possible and these are, as in the OAIS standard, indicated by initial capitals, e.g. Ingest. As in OAIS we use the term Archive to denote any organization devoted to long-term preservation.

In the remainder of this article we present the results of the second phase of the project describing cost aspect of Ingest and in particular cost associated with normalization: In section 2 we present our analysis of the Ingest functions and the identified cost dependencies. In section 3 we analyze format obsolescence and different cost drivers in digital migration, including migration quality, timing and frequency. In section 4 we describe how the costs of migrations have been modeled in the CMDP, including the cost of monitoring and executing migration actions. We describe the results of testing the CMDP on empirical cost data in section 5, and conclude in section 6.

2. INGEST OF DIGITAL INFORMATION

As a first step in identifying activities related to the Ingest of digital information into an Archive, a flow diagram was prepared based on an analysis of the functional descriptions in the OAIS standard (see Figure 1). Note that the activity Generate SIP (Submission Information Package) is not part of the standard (see explanation below). The flow analysis also helped avoiding that critical activities were overlooked or accounted for more times.

In addition to the OAIS standard we consulted the PAIMAS standard. The strength of PAIMAS is that it includes a checklist for defining a Submission Agreement, specifying all the details about a submission necessary for ensuring long-term preservation of the information. PAIMAS also describes activities related to the transfer of data and the validation of the transfer.

2.1 Submission Projects

PAIMAS divides a submission project in four phases:

1. The purpose of the preliminary phase is to determine whether a submission project is feasible and financially viable. The phase comprises the first contact between Producer and Archive, the provisional definition of the project's objective and context, a draft description of the digital information and its structure, and the writing of a draft Submission Agreement.
2. The formal definition phase negotiates the Submission Agreement between the Producer and the Archive. It describes the design of the SIP and the digital information to be submitted. Also it determines legal and contractual terms as well as security, and describes how transfer and validation of the transfer are to take place. Finally, it sets up a timeframe for the project.
3. The transfer phase ensures that the SIP is transferred from Producer to Archive, and that the Archive's initial processing of the information takes place according to the Submission Agreement.
4. The purpose of the validation phase is to ensure that the transfer of the digital information is validated according to the requirements outlined in the Submission Agreement.

Definition of a formal Submission Agreement does not necessarily occur as part of a submission. This depends on the nature of the submission and the power balance between the Producer and the Archive. In some countries an Archival Act can mandate archives to specify SIP designs, and in this case the balance of power is in favor of the Archive. In other scenarios, the Archive has to accept SIPs from the Producer as they are. This is typically the case within the library and research sector.

Even if no Submission Agreement is formally required it may still be important for the Archive to analyze the PAIMAS checklist for the Submission Agreement and determine how these issues will be handled. As such, the Submission Agreement constitutes an important part of any Archive's policy and strategic planning documentation.

2.2 Ingest Flow and Cost Dependencies

Below we describe activities under Ingest in detail and the identified cost dependencies. No specific costs of are reported in the article since they often depend on several preconditions, such as type of material, volume and format complexity. To calculate actual costs please consult the spreadsheet.

The cost of the core preservation system, i.e. the system, which e.g. manages notifications and the reception and transfer of information, is assumed to be accounted for in the Common Services functional entity of CMDP. This module has however not yet been modeled in CMDP.

2.2.1 Negotiate Submission Agreement

In the OAIS Model the Submission Agreement is negotiated by the Producer and the function Negotiate Submission Agreement under Administration. The agreement must cover all parts of the submission project, including a data submission schedule and an assessment of the required resources to support the submission.

The costs of negotiating a Submission Agreement are first and foremost dependent on the balance of power between the Producer and the Archive, the diversity and complexity of the data, how well the data are documented, and the size of the submission project.

2.2.2 Generate SIP

If an Archive bases its preservation strategy on migration and receives SIPs in production formats, which are not regarded as suitable for long-term preservation, it is common practice to

[1] www.costmodelfordigitalpreservation.dk

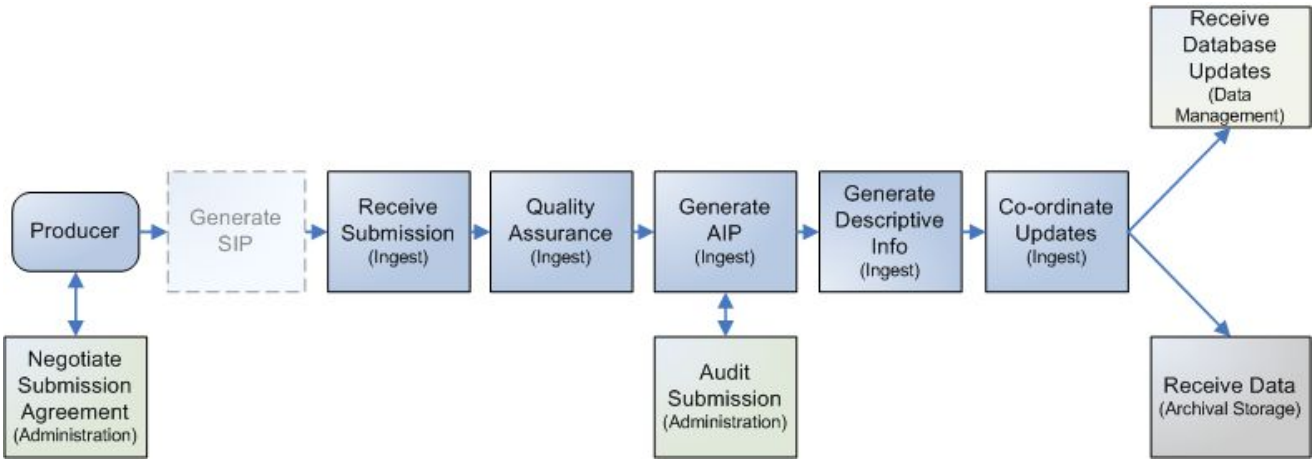


Figure 1 The flow between the OAIS functions from Producer to Archive on submission of digital records

normalize the information at Ingest, i.e. to migrate it from production to preservation formats and structures. Often normalizations are more complex and thus more costly than migrations from one preservation format to another. In OAIS, normalization is carried out by the Ingest function Generate AIP (Archival Information Package).

However, if the balance of power is in favor of the Archive, it can require the Producer to submit SIPs with normalized data in validated submission formats and/or enrich the packages with metadata before the SIPs are transferred to the Archive. As has also been noted as part of the five year review process of the OAIS standard [8], such preparation of the SIPs is not explicitly accounted for in the OAIS Model. In order to be able to calculate these costs, we have added the optional function Generate SIP to the CMDP.

If Ingest includes normalization this entails considerable costs, irrespective of whether these costs are carried out by the Producer or the Archive. It is the balance of power between the Archive and the Producer that determines who pays for the cost of normalizations. The cost of normalization and dependencies are described in detail in section 4.2.

2.2.3 Receive Submission

When the Submission Agreement has been concluded, the Producer transfers the SIPs to the Archive, where they are received by the Ingest function Receive Submission and placed in temporary storage. The transfer may be by movable storage media such as DVD or hard disk, or via a network.

In the CMDP we assume that reception of SIPs is an automated process, and thus not cost critical. We have also excluded costs of providing temporary storage for receiving SIPs because this extra storage capacity is likely to be reused for other activities.

2.2.4 Quality Assurance

The SIPs are then checked for errors by the Quality Assurance function, typically by a check-sum control. If the packages are in order a confirmation of reception is sent to the Producer. If there are errors the Producer is informed, so that the packages can be corrected and transferred once again. It is important to notice that in the OAIS Model this function only verifies the integrity of the data. It checks neither the authenticity nor the intellectual content, which in OAIS is managed by the Audit Submission function under Administration (see section 2.2.6).

This quality assurance process is assumed to be automatic and thus only entails costs for the establishment and maintenance of the quality assurance system as well as the potential error handling.

2.2.5 Generate AIP

The Generate AIP function transforms SIPs to AIPs, and may entail normalization. The function may also request that the functional entity Data Management provides additional information necessary for a full description of the package.

As the costs of generating, SIPs those of generating AIPs are described in detail in section 4.2.

2.2.6 Audit Submission

The Audit Submission function is part of the functional entity Administration. It checks whether the generated AIPs fulfills requirements and sends an audit report back to the Generate AIP function. If any errors or defects are identified the Producer is noticed and can then transfer the SIPs again. The validation phase specified in the PAIMAS standard corresponds to the Audit Submission function as well as the Quality Assurance. Audit typically comprises ensuring that the information packages are complete, that integrity is maintained, and that they fit the data model, including that the agreed data formats have been used and their syntax have been maintained.

Depending on the requirements defined in the Submission Agreement, the audit of AIPs can be cost intensive, because they are often manual.

2.2.7 Generate Descriptive Information

The Generate Descriptive Information function subsequently extracts Descriptive Information, i.e. primarily metadata used to search and retrieve the packages, from the AIP and other related sources, and sends the information, via the function Co-ordinate Updates, to Data Management.

The extraction of Descriptive Information is assumed to be an automatic process, and therefore not cost critical. Note that institutions may use considerable resources for providing metadata to the objects at Ingest. OAIS does not explicitly include such data qualification, and therefore it may be accounted for under Generate SIP in the CMDP.

2.2.8 Co-ordinate Updates

The Co-ordinate Updates function then sends the AIPs to Archival Storage, which confirms receipt and assigns an ID for the AIP package when storage has been executed and verified. Co-ordinate Updates includes this ID in the Descriptive Information and sends it on to Data Management.

The co-ordination with Data Management and transfer to Archival Storage is assumed to take place automatically and these activities are therefore not regarded as cost-critical in the CMDP.

3. COST DRIVERS IN MIGRATION

Below we analyze format and system obsolescence and its influence on when migrations should take place. Thereafter we describe three important drivers of cost in digital migration actions, namely migration quality, timing and frequency.

3.1 Format and System Obsolescence

Virtually irrespective of its format data in themselves are of little value, as they require a system, which can interpret the data format in order to reproduce the data content in an understandable form. The obsolescence of formats is therefore dependent on the obsolescence of the software that is to interpret the format.

With regard to data preservation, it is not sufficient that a program can interpret data in their current form (input format), as it must also be able to write data in a suitable contemporary format (output format). Especially in earlier times it has been emphasized that there was often a viewer for a format, which was therefore not obsolete. This is not a tenable argument; however, as the result is dependence on new generations of the viewer, and in addition the data cannot be further processed in other systems. The answer thus merely leads to the new question of when the systems in question that can read data in its present format (input format), and write the data in a contemporary output format, are obsolete?

For as long as new generations of systems are developed that can read data in its present format, and write the data in a contemporary output format, there is no real problem of obsolescence. This does require, however, that the system is tested, and that the reproduction is acceptable. When a generation of the system is developed in which the data format in question can no longer be interpreted, or just cannot be written in a suitable contemporary format, it is necessary to use the previous generation of the system to do this, thereby becoming dependent on its lifetime.

More and more formats can be interpreted by systems that are several generations younger, although naturally there are limits. It is therefore necessary to use the previous generation of the program to read the formats and save them in a contemporary output format. As for most new generations of programs and data formats there are a number of functionalities and derived data that are not supported in the newest generation.

The lifetime of systems does not end on the same day that a new generation of the system is born, or a competing system takes over the market. The lifetime of systems is dependent on the costs of their use and maintenance. For as long as a system can run on contemporary hardware and be integrated with contemporary systems the costs of its use and maintenance are manageable. Thus, neither the system nor the data formats it can interpret and write to are obsolete.

A known example of obsolescence is the BBC Domesday Project: In 1986 the BBC published an extensive modern multimedia edition of the famous Domesday Book that describes England in the 11th century. The BBC Domesday edition consisted of letters, maps, images, statistical data, videos, etc., stored on two interactive laser discs, LaserVision Read Only Memory (LV-ROM). In 2002, it was feared that the discs would become unreadable due to the technological obsolescence of the data storage medium and it was necessary to use migration, emulation and re-digitization in order to preserve the data. This was technically possible with great difficulty, and the high costs were a clear indication that the formats had become obsolete.

“The lesson of this digital preservation project is that if you have enough time, individual skill, dedication and imagination then almost anything is possible, provided that you don't leave it too late. If you start counting the cost this may seem an expensive project, but then the value of the record is high too - and that applies equally to the original Domesday Project. There is of course a great need to preserve other electronic records in a routine and predictable manner, and this rescue project is not a suitable model to be followed in such cases. The National Archives is working on ways to make this possible in future” [6].

This is despite the fact that from the outset the project’s creators were aware of the preservation risk and had in due time submitted data and documentation to an archive that did not handle the matter satisfactorily.

“The deputy editor of the Domesday Project, Mike Tibbets, has criticized the UK's National Data Archive to which the archive material was originally entrusted, arguing that the creators knew that the technology would be short lived but that the archivists had failed to preserve the records effectively [16].

Do we always have to rely on existing systems to be able to read data in a given format? In practice yes, since even with exhaustive documentation of the format it is normally a very demanding task to develop a system, to read data in one format and write it in another. The exception is the very simple formats for which, at a modest cost, it is possible to develop systems that can read data in one format and write it in another. Examples include TIFF, UTF-8 or XHTML.

3.2 Migration Quality

As for many other costs, the quality level of migrations is decisive to the level of costs. Migration quality is determined primarily by the choice of the output format, and by the error tolerance on migration of data from the input format to the output format.

3.2.1 Selection of Output Format

High quality in terms of an advanced output format, which enables preservation of a wide range of functionalities, rather than a simpler output format will entail significantly higher costs. This is because on migrating data from input format to output format programs must handle how all data in the input format is migrated

to an equivalent place in the output format, and it must be controlled that this has taken place (see below). For example, migration from one word processing format to another word processing format will result in higher costs than migration to a simple format in the form of a graphic bitmap format, as the word processing format contains far more information than a graphic format. This is a general observation, since in practice the situation may be that the system that migrates data from the input format to an advanced output format is far superior to the system that migrates data to a simple output format. The choice of output format is also essential to determining how *often* migration should be performed (see section 3.3).

3.2.2 Selection of Error Tolerance

With regard to error tolerance on migration of data, high quality in the form of a low error tolerance will bring about significantly higher costs than a high error tolerance. This is because a low error tolerance will typically require extra funds for the provision, operation and further development of the system for the migration. In addition, it will be necessary to use extra resources for error control, and especially error correction. Irrespective of the choice of error tolerance there will normally be higher costs for the error handling of an advanced output format than of a simple output format. This is because there is more chance of something going wrong, and it is more expensive to correct the individual errors.

Selection of output format and error tolerance can furthermore be combined, depending on the purpose of preservation and the data content. Note that an advanced output format thus does not necessarily entail a low error tolerance, just as a simple output format does not necessitate a high error tolerance.

3.3 Migration Timing

An important factor with regard to the costs of migration is when in the archival lifecycle, migration should be performed. There are different tactics for when it is best and least expensive to migrate, including to which output format.

3.3.1 Migration to Standardized Format

One tactic is to migrate data to a contemporary standardized format, as seldom as possible. The argument behind this tactic is that by migrating to a contemporary standardized format the number of migrations is reduced, and thereby the risk of unintended changes. The reason is that the lifetime of a standardized format is expected to be significantly longer than for other formats, as several systems will be able to read data in the format and write it in another. In addition, the standardized format should make it less expensive to provide, operate and maintain systems for actual migration, due to the larger supply available.

On the other hand, the number, market penetration and system support of contemporary standardized formats is estimated to be modest. It is therefore necessary to either select simple output formats, or to perform migration almost as frequently as if the next generation of the input format had been chosen as output format.

3.3.2 Migration to the Latest Format

Another tactic is to continuously migrate data to the most recent output format. The argument behind this tactic is that it adopts the situation of other IT users with a need to migrate data from the previous generation of the format to the latest as correctly and inexpensively as possible. This makes it possible to benefit from the systems for the latest generation of the format, which must be assumed to be the best for reading the immediately preceding generation of the format.

On the other hand, the frequent migrations are cost intensive and increase the risk of unintended changes. Moreover, the programs for the newest generation of the format are not always the best to interpret the previous generation [10]. Sometimes it is necessary to wait for the following generation to achieve better reproduction. In addition, suppliers and users generally seem more interested in creating new data in new formats, rather than reading older data in older formats in the new generations of the systems, that nothing particular is done to facilitate migration. It is thus difficult to find systems that handle mass migration of the previous to the current generation of the format.

3.3.3 Migration on Demand

A third tactic is called migration on demand and entails that if the data are in a relatively common and documented format the data are retained in the original format and not migrated to another format until the data are requested. The argument behind this tactic is that it is estimated that the number and variation in the use of data formats is continuously narrowing, and that market penetration, openness and documentation are widening. The probability that in a few years it will be possible to read a previously relatively common and documented format is therefore so high that there is no reason to perform migration before then. This saves a large number of intervening cost intensive and hazardous migrations.

On the other hand, the risk is considered by some to be too high, i.e. the probability that after a number of years there will, after all, not be any system that could interpret the format. In addition, depending on the output format, it is often an advantage to migrate shortly after the data are created, as many formats are not isolated, but depend on external data, for example fonts in the system, or references to images or other data outside the format that may have been altered after a number of years. These are external dependencies of which the encapsulation requires systems that have to be acquired, operated and further developed. Some standard programs, such as MS Word 2010, now support partly embedded fonts.

3.4 Migration Frequency

The immediate answer to how often migrations should take place is as seldom as possible, while bearing in mind the risk of obsolete data. This is because each migration entails a risk of losing information when data are migrated from one format to another, and because each migration entails costs.

On the basis of the current situation our tentative estimate of when a format is obsolete is eight to 20 years after its introduction on the market.

Twenty years is based on the furthest horizon we dare estimate within digital preservation. Eight years is based on the time within which we estimate that it will generally still be possible to run a program that can read data in its input format and write it in a suitable, contemporary output format.

3.4.1 Format Lifetime Parameters

It is extremely difficult to estimate the lifetime for a given format between the extremes of eight and 20 years, but we assess the vital parameters to be market penetration, complexity and documentation of the format. Lifetime increases with widespread

use, low complexity and good documentation. The three parameters are mutually dependent, which does not make the estimate easier. Simple, well-documented formats are often widely used, and simple formats are often well-documented.

In this context market penetration concerns the number of users, but especially the number of different systems that use the format. IT is a market with considerable network effects, and the aim is to develop programs that can fully read a competitor's format, but only write in their own formats; otherwise it is necessary to compete on the competitor's home turf, or on an equal footing.

Complexity is dependent on the number of types of information in the format, including the functionality in the system that is reflected in the format. Highly complex formats are often replaced more quickly (than formats of low complexity) by new generations of the format, as producers or users require even more functionalities. As stated, formats of very low complexity can be independent of existing systems because on the basis of the documentation, if it is good, it will be possible, without prohibitive costs to develop a system to interpret the format.

Documentation concerns the description of the structure and use of the format. A characteristic of good documentation is that it gives others besides the original creator of the format a feasible opportunity to develop systems that can interpret the format. It will at times also be necessary to have partial documentation of the system in order to understand how to interpret the format. For documentation to be good it must first of all be accessible, and secondly include the entire structure and use of the format, and finally be explanatory, i.e. intended to ensure that others besides the original developers can understand the format.

4. COST OF MIGRATION ACTIONS

There are numerous costs related to migration, of which the most important are the provision, operation and further development of systems for:

- Ongoing monitoring of which formats are obsolete, and of which the content must be migrated to other formats.
- Actual migration of data from one format to another, including control that the data is not changed unintentionally.

The following cost-estimates are based on own experience and a review of the literature on this subject [1]. With regard to the further development of the migration cost formula, we have been inspired in particular by the guide: Software Development Cost Estimating Guidebook [14].

4.1 Costs of Monitoring

Costs must be defrayed for the provision, operation and further development of a system for identification and registration of all formats for all data, stating the precise version of each data unit.

In practice this entails that on ingest of data in the preservation system all data are analyzed, so that its formats can be identified and registered, and so that all data in a given format can be retrieved when it is transferred to another format.

This task can be handled by the Producer if the Archive can get the Producer to undertake the task, and trust the result, but in practice most preservation institutions will handle this themselves. Identification should in practice be followed by validation and partial characterization. This is because far too much data does not comply with its format, and that many formats are so rich in content that it can be necessary to have information on their characteristics, i.e. which parts of the format contain data.

4.1.1 Provision of Monitoring System

Provision of such a system currently requires that it has to be developed, although there are partial solutions in the form of JHOVE2[2], PRONOM and DROID[3]. We estimate that the costs of provision of the core of a modular system that via specific modules for the individual formats can perform reasonable identification, partial validation and a small degree of characterization will be 12-24 person months.

The costs of the development of the individual modules depend on the formats' complexity and documentation, and are estimated to be respectively exponentially increasing and diminishing. We estimate that the cost per format will be from a few person weeks for simple formats to several person weeks for advanced formats.

Going beyond what we unclearly call reasonable identification, partial validation and a small degree of characterization, we estimate that there will be a highly exponential increase in the costs. It has, for example, still not been possible to achieve a complete validation of PDF/A. It is currently necessary to use validators from several suppliers to cover as many areas as possible. It will not be possible to avoid incorrect identification or incorrectly formatted data. In practice, it must be hoped that the programs to migrate data to other formats are relatively error tolerant. It will not be possible to avoid a few errors without very high costs.

4.1.2 Operation of Monitoring System

We assume that monitoring takes place by manual review of the list of formats used and comparison of their development in the market, in order to assess whether some formats are becoming obsolete. Work is also taking place on the establishment of a joint international format register, the Unified Digital Format Registry (UDFR)[4], which will be able to streamline monitoring. Monitoring of the market means that for each format there is one or several system(s) that must be registered and stated as necessary to interpret the format. These systems' lifetimes must also be assessed, including whether the format is supported in the newest generation of the system.

The task of monitoring is highly manual, and we estimate that the cost is proportional to the complexity of the format. On this basis it is estimated that monitoring will take from a few person days to a few person weeks, and that it will most frequently have to take place every second year for a given format.

4.1.3 Maintenance of Monitoring System

Besides general maintenance, the maintenance of the system, for example in connection with a new operating system, also includes the development of new profiles for identification, validation and characterization of any new formats that the Archive might use.

[2] https://bitbucket.org/jhove2/main/wiki/Home
[3] www.nationalarchives.gov.uk/PRONOM/Default.aspx
[4] www.udfr.org/

4.2 Costs of Migration

In terms of costs the migration of data from one format to another can be divided into provision, operation and maintenance of migration systems:

4.2.1 Provision of Migration System

We assume that a migration system has the following modules to handle the required tasks:

A general module that on the basis of central registration of data and their format can retrieve the data in an information package (a SIP or AIP) of which the format is estimated to be obsolete, and unpack this data.

A general module to manage all information packages and data retrieved in the obsolete formats, as well as their status, throughout the migration process. For each body of data in a format the module must request the specific module created for each format, register the result, and if successful send the migrated data in its new format for repackaging with the unaltered data from the package, so as to create new packages. To ensure efficiency the module must be able to parallelize its requests.

Specific modules for each format that ensure that the data in the format is migrated with the system considered to be the most suitable for the process and in the required quality. These programs will normally be the same as were registered in conjunction with the monitoring of the format's obsolescence. To be able to automate migration the module must be able to control parts of the program's behaviour, for example so that it is not stopped by enquiries from the program. If an advanced output format is selected there may also be a need for further management of the program in order to migrate all the required information to the output format.

The costs of developing the above system are considerable, and reuse of others' solutions is an obvious alternative. We do not know any turnkey solutions, but a number of sub-solutions, such as Apache Hadoop[5] or Berkeley Boinc[6], might be used.

We estimate that development of the general modules takes 12-24 person months. The costs of the specific modules are not necessarily proportional to the number of formats, if a series of formats use the same program for migration. The test of the correct functioning of the module with a given format is, however, proportional, and the cost can therefore be almost proportional. Reuse of others' solutions is an obvious path to take, but we do not know of any such solutions. For each format, primarily the advanced formats, where there is a need, there are often full or partial solutions, such as Apache POI or Microsoft Open XML Format SDK[7], that can manipulate the running of a program or directly access the format. We assess, however, that the cost of directly accessing the format in the case of advanced formats, such as ODF or OOXML, exceed what is feasible for an individual preservation institution. The institutions must therefore await development in a wider community if the quality is to exceed that offered by turnkey programs.

We estimate the cost per format to be exponential to the format's complexity, and vice versa in terms of error tolerance. Furthermore, we estimate that the development of a module for a simple format with a low error tolerance will take a few person weeks, while an advanced format with a low error tolerance will take several person weeks.

[5] http://hadoop.apache.org/
[6] http://boinc.berkeley.edu/
[7] http://msdn.microsoft.com/en-us/library/bb448854.aspx

4.2.2 Operation of Migration System

The costs of operating the system are primarily related to error handling, which depends on how reliably the system has been developed to operate. In this respect the costs of development and subsequent error handling are often inversely proportional, and it is not easy to calculate the optimum distribution.

Error handling comprises actual operational interruptions in the areas for which the system has not been developed to operate reliably enough. It also includes the identification of errors in the individual modules, when a format cannot be migrated as expected. Finally, error handling concerns errors that the system does not know that it makes, and which can only be detected via subsequent random sampling. In other words, handling errors that it is known will arise; errors that are assumed to arise; and errors that are not expected to arise. When the errors have been identified it is necessary to decide whether they are to be corrected, and if so, how.

Depending on the migration quality selected, primarily the complexity of the output format and the migration's error tolerance, we estimate that monitoring per format per TB (Terabyte) takes from one person day to a few person weeks. Furthermore, we estimate that error correction takes up to ten times longer than monitoring.

Even though the costs can be compiled per format, there are still economies from migrating several formats simultaneously, for example on packaging and unpackaging, storage, and error handling. As formats do not die on the same day that they are declared to be obsolete, several obsolete or virtually obsolete best practice is to gather formats for simultaneous migration.

4.2.3 Maintenance of Migration System

Maintenance of the system comprises general maintenance, for example in connection with a new operating system, and the development of new modules for new formats.

5. TEST OF CMDP ON COST DATA

A questionnaire was sent to a number of public Danish authorities in order to collect information on their actual consumption of time and resources to produce information packages of data from IT systems in connection with submission to The Danish National Archives. The data collected have been used to test and adjust the Ingest module in the CMDP. If the authorities used an external supplier to prepare the information package, cf. Generate SIP, they were also requested to submit a copy of the contract for the assignment in order to obtain a full overview of the costs.

The questionnaire was sent to 34 authorities, of which approximately half replied. The responses received point in many directions and show that the authorities found it difficult to understand the questionnaire and compile the consumption of resources. Based on the responses received a tentative conclusion for large submission projects (>160 person hours) is that project management costs approximately 13% of the total submission project. The identification and the description of the digital objects and their references accounts for approximately 16%. Normalization accounts for approximately 66% and the testing of

the information package accounts for approximately 5%. The responses concerning the time spent on the physical submission are not included in the study as the responses showed that the question was not understood correctly. Furthermore, in the case that the authorities have used consultants, a high price is not necessarily equivalent to high earnings for the consultant, as the price/earnings ratio is not equal for all consultants.

6. CONCLUSIONS

In overall terms, we believe that the developed method of identifying Ingest costs is viable, although the cost model is not yet sufficiently detailed to give accurate results for all types of records: All empirical data originates from ingest of archival records, which means that the model is currently best suited to estimate the costs of this particular type of material.

An important conclusion from the survey on submission projects was that the normalization of formats is by far the highest Ingest cost, namely around two thirds of the total costs. The fact that normalization also entails cost-sensitive choices such as migration frequency, timing, quality, error tolerance, emphasizes that this particular cost requires very special focus when considering the precision of the cost model. Likewise, the study indicates that the balance of power between Producer and Archive has great influence on the costs, and their distribution, so that this is an essential parameter in the model.

The study also confirmed a former key finding: The choice of the digital object (the format), its complexity and volume as the basic calculation units, makes the model potentially generic and thereby capable of calculating the costs for various digital collections. In order to achieve accurate results for all types of digital materials there is, however, a need to expand with several parameters for each object type, for example number of objects.

Likewise, more work is required to increase the precision of the model. By default the CMDP has a number of estimates such as format complexity, lifetime and thereby migration frequency, which are dependent on the actual preservation scenario, highly uncertain and subject to debate. In order to address this problem, the model makes it possible to state other values than those proposed default.

Generally, our work shows that preservation institutions depend to a great degree on being able to use standardized solutions, as it would be very expensive for them to develop a number of tailored tools corresponding to the number of types of ingested data.

The implementation of the model in the spreadsheet has proved to be problematic. The requirements of transparency and precision cannot be fulfilled simultaneously. In a new version of the model, with greater precision, it will therefore probably be necessary to sacrifice some of the immediate transparency and state the formula in code. The lack of an actual user guide and user interface to the spreadsheet is another deficiency, as the model in its current form is very difficult for external parties to use.

Currently, we have funds for 1½ man-month for developing the cost module for Archival Storage, which is a somewhat easier task as we have longer experience with these functions and more empirical data is available.

If the precision of the CMDP is to be increased, the remaining modules of CMDP are to be developed, and the model expanded to account for different preservation strategies and different digital collections, additional work and funding is needed. For the purpose of further development of the model, the project has stayed abreast of the international development of economic models for digital preservation. It is our hope that this focus will lead to formal or informal cooperation with other stakeholders in the future, as it is assessed that both the interest in and the necessity of greater certainty in this field are generally considered to be substantial.

7. ACKNOWLEDGMENTS

Thanks to the Danish Ministry of Culture for funding this study.

8. REFERENCES

[1] Beagrie, N., Lavoie, B., Woollard, M., 2010. Keeping Research Data Safe 2, Final Report, Charles Beagrie Limited, www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf.

[2] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 2010. Sustainable Economics for at Digital Plant: Ensuring Long-Term Access to Digital Information, Final report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

[3] Consultative Committee for Space Data Systems (CCSDS). 2002. Reference Model for an Open Archival Information System (OAIS), 650.0-B-1, Blue Book (ISO14721:2003). http://public.ccsds.org/publications/archive/650x0b1.pdf

[4] Consultative Committee for Space Data Systems (CCSDS). 2004. Producer-Archive Interface Methodology Abstract Standard (PAIMAS), CCSDS 651.0-M-1, Magenta Book. http://public.ccsds.org/publications/archive/651x0m1.pdf

[5] Cooper, R., Kaplan, R.S., Maisel, L.S., Morrissey, E. & Oehm, R.M. 1992. Implementing Activity-Based Cost Management: Moving from Analysis to Action. New Jersey. Montvale, Institute of Management Accountants.

[6] Darlington, J., Finney, A. & Pearce, A.. 2003. Domesday Redux: The rescue of the BBC Domesday Project videodiscs, Ariadne Issue 36. www.ariadne.ac.uk/issue36/tna/intro.html

[7] Data Archiving and Networked Services (DANS). Costs of Digital Archiving vol. 2. www.dans.knaw.nl/en/content/categorieen/projecten/costs-digital-archiving-vol-2

[8] Higgins, S. & Boyle, F. 2006. Response to CCSDS's comments on the OAIS Five-year review: recommendations for update, The Digital Curation Centre (DCC) and The Digital Preservation Coalition (DPC). www.dpconline.org/events/previous-events/427-oais-5-year-review-follow-up

[9] Hole, B., Lin, L., McCann, P. & Wheatley, P. 2010. LIFE3: A Predictive Costing Tool for Digital Collections, In: Proceedings of iPRES 2010, 7th International Conference on Preservation of Digital Objects, Austria www.ifs.tuwien.ac.at/dp/ipres2010/papers/hole-64.pdf

[10] Karjalainen, M. 2010. Large-scale migration to an open source office suite: An innovation adoption study in Finland, Academic Dissertation, Faculty of Information Sciences of the University of Tampere. http://acta.uta.fi/pdf/978-951-44-8216-8.pdf

[11] NASA Cost Estimation Toolkit (CET). http://opensource.gsfc.nasa.gov/projects/CET/CET.php

[12] Kejser, U.B, Nielsen, A.B., Thirifays, A. 2011. Cost Model for Digital Preservation:Cost of Digital Migration. In: The International Journal of Digital Curation, Issue 1, Vol. 6, pp. 255-267. www.ijdc.net/index.php/ijdc/article/view/177

[13] OECD. 2004. International Standard Cost Model Manual to reduce administrative burdens. www.oecd.org/dataoecd/32/54/34227698.pdf

[14] Software Technology Support Center (STSC) Cost Analysis Group, U.S. Air Force. 2010. Software Development Cost Estimating Guidebook www.stsc.hill.af.mil/consulting/sw_estimation/SoftwareGuidebook2010.pdf

[15] Slats, J. and Verdegem, R.. 2005. Cost Model for Digital Preservation. Proceedings of the IVth triennieal conference, DLM Forum, Archive, Records and Information Management in Europe. http://dlmforum.typepad.com/Paper_RemcoVerdegem_and_JS_CostModelfordigitalpreservation.pdf.

[16] Tibbets, Mike, 2008, ACM Committee on Computers and Public Policy, Forum on Risks to the Public in Computers and Related Systems, Vol. 25: Issue 44. http://catless.ncl.ac.uk/Risks/25.44.html#subj

The Costs and Economics of Preservation

Neil Grindley
JISC
Brettenham House, 5 Lancaster Place
London, WC2E 7EN
+44(0)2030066059
n.grindley@jisc.ac.uk

ABSTRACT
Given that preservation is now a fairly well-described problem, it should, in theory, be possible to calculate with a reasonable degree of accuracy what costs are likely to accrue to an organisation that has responsibility for the long-term stewardship of digital assets. This paper will introduce and describe some of the work that has been carried out over the last 5 years to help institutions and research groups to understand both the cost and the economics of preservation, and to examine the difference between those concepts. It will also describe ongoing phases of work that are being funded in the UK by JISC that are attempting to further advance understanding in this area and where possible apply or implement previously theoretical approaches. Some indication will also be given as to where collective international effort may be of universal benefit.

Keywords
Preservation, costs, economics, models

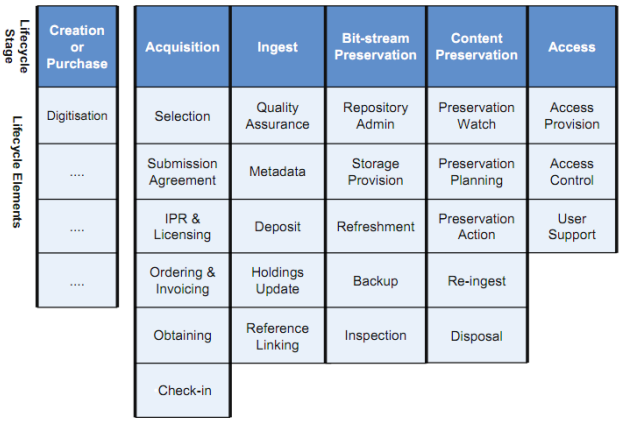
1. INTRODUCTION
In the last five years, some groundbreaking work has been done relating to the costs and economics of digital preservation. The LIFE project[1] undertaken by University College London and the British Library devised and refined a lifecycle costing model for digital objects which incorporates a generic preservation cost component and a costing tool. The Keeping Research Data Safe (KRDS) project[2] examined this same issue but specifically with a focus on the long-term management of research data. In the US (with some UK involvement) the Blue Ribbon Task Force on Sustainable Digital Preservation and Access[3] (BRTF) spent two years analysing the economic conditions under which a variety of digital object types might best be maintained for future utility.

The purpose of this paper is to look at the various different ways that these three initiatives are currently being followed up and to propose future actions and reactions in response to them. The three follow-on activities are all being funded by JISC in the UK and have not yet been widely disseminated or discussed, either in the UK or internationally.

iPRES2011, Nov. 1–4, 2011, Singapore.

2. THE LIFE PROJECT
The LIFE project (Lifecycle Information for e-Literature) began in 2005 and consisted of three phases of work to investigate the possibility of defining an entire lifecycle model for a digital object and to then relate the parts of that lifecycle to the likely management and maintenance costs that might be incurred by the owners or keepers of the digital asset in question. The original context of this work and the initial focus was on estimating the cost of large homogeneous collections of materials, such as might be looked after by a national library or a large research intensive university. As such, the resultant model and tool may be more suited to a certain types of collections management procedures rather than others. Despite any perceived limitations, however, it is clear that the wider community valued this attempt to estimate retrospective and future costs and the LIFE project final reports from all three phases of work have been extensively downloaded and referenced.[4]

Figure 1 The LIFE model[5] (c.2010)



The influence of the project can also be measured by follow-on work and the Danish National Library and Archive have used the LIFE project model (in a somewhat adapted form) for their own purposes.[6] The LIFE-SHARE project[7] in the UK, based at University of Leeds, has also picked up on the LIFE modeling work and has used it to investigate the skills and strategies required for managing end-to-end digitization processes, including preservation of the created content.

At the end of phase 3 of the LIFE project, a functional tool was produced, based on a series of complex Microsoft Excel spreadsheets, that meant that users could input parameters and figures into a form, and then view costs over various timescales

and with various degrees of precision (depending on the level of detail of the original parameters) to support decision making about the cost of storing and managing digital materials over time.

2.1 Piloting the LIFE Model

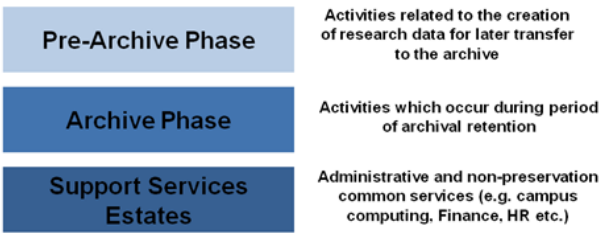
To follow up and properly exploit the 3 phases of work on the LIFE model and the tool, JISC commissioned HATII (University of Glasgow), working under the auspices of the Digital Curation Centre, to take the LIFE tool out into the UK HEI community in order to check whether a tool that was initially devised and developed with large document-type collections in mind could be applied in the university context where a different scale and scope of materials might require analysis, and where different input data (in particular relating to salaries and overheads) may be apparent.

As stated by the project page on the DCC website, it is anticipated that "the participating HEI repositories will benefit from a greater understanding of their day to day running costs and may even be able to identify inefficiencies in their current processes. In the longer term, this increased understanding of actual costs may inform strategic planning and policy development at the institution. The cost data provided by the targeted repositories, once anonymised, will have the potential to enrich the LIFE model for all subsequent users of the tool and will help to provide more accurate cost estimates for a broader group of organisation types."[8] For the conference presentation itself, it will be possible to give a summary of the results of this short pilot phase as it is scheduled to deliver the final report around about the time of writing this submission. One of the purposes of presenting this work to the IPRES audience will be to elicit support for further engagement with the tool and to create opportunities for further international collaborative work around the topic of cost modeling.

3. KEEPING RESEARCH DATA SAFE (KRDS)

The first phase of KRDS[9] took place in 2007 and was one of the early pieces of work that JISC commissioned in the area of research data management. Subsequent to this JISC established a substantial progamme of work[10] to support this activity but it was clearly seen from the outset that the approach taken with the LIFE project could usefully be extended to cover information defined as 'research data', and that the challenges associated with this form of information were different and discreet enough to require separate investigation.

Figure 2 Highest Level of the KRDS Activity Model[11]



The KRDS work was led by Neil Beagrie (Charles Beagrie Ltd.) and was carried out in collaboration with partners including OCLC and the UK Data Archive (see KRDS web page for the full list of contributors and partners). Following two phases of KRDS project work and some further funded activity to produce discreetly bundled related material (i.e. a fact sheet; a user guide; and detailed and summary activity models), an additional collaborative activity occurred in conjunction with an existing project called I2S2 (Infrastructure for Integration in Structural Sciences), based at the University of Bath. The objective of the I2S2/KRDS project was "to test, review and promote combined use of the Keeping Research Data Safe (KRDS) Benefits Framework and the I2S2 Value Chain Analysis tools for assessing the benefits of digital preservation of research data." This collaborative work makes more explicit the work relating to the benefits (as well as the costs) of managing research data that KRDS project began to seriously address in its second phase, and which the original I2S2 project engaged with at the outset as part of proving the value of integrated research infrastructure.

3.1 The Costs Observatory

In the course of presenting the conclusions from the second phase of the KRDS work, it was suggested to JISC (by Neil Beagrie) that some consideration should be given to the establishment of a 'costs observatory' that would facilitate the gathering, processing, analysis and dissemination of appropriate costs information relating to the management of long-lived data. The motivation for this suggestion originated from the experience of trying to collect authentic, useful and comparable cost information. It proved to be an extremely challenging task, particularly devising ways of comparing the data across different types of organisations, and one of the most prominent conclusions was that it would be far easier and more effective to setup a method of capturing cost data going forward than to try and retrofit comparison schema to diverse information sources.

To examine if the concept of a costs observatory was a workable idea, JISC commissioned a short (10 week) consultation and scoping study from Key Perspectives Ltd. during the period May – July 2011. According to the text of the invitation to tender, the "principal target outcome [of the currently imagined 'costs observatory'] would be to influence strategic planning and policy formation within institutions and enable them to make wiser, more realistic and cost effective decisions about managing information."[12] The detailed objectives (of the proposed observatory) were to:

- pro-actively seek and collect costs information relating to the short, medium and long term management of digital materials and data
- develop capability and status as a trusted broker of sensitive and confidential financial information
- analyse the financial data and produce reports and recommendations for universities and colleges (HEI's), funding bodies and strategic agencies on issues to do with the costs and economics of managing information
- support the UK HE sector with determining its existing and predicted Information management costs
- monitor and identify relevant economic, legislative and environmental issues
- liaise and co-ordinate with relevant service and information providers

Key Perspectives Ltd. did some analysis and scenario-building work and consulted with various representatives from the UK HE community on the efficacy of the proposed 'costs observatory', and then presented their conclusions to JISC in a report. In relation to one of the principal concerns laid out in the ITT, i.e. the scope of data to be collected – or to put it another way – the type of information (e.g. research data, administrative information, systems data, student records, learning and teaching materials, etc) that the observatory would gather, the report concluded that the focus would sensibly be on research data. This conclusion was arrived at through a combination of logistical possibility; declared community requirement; most pressing urgency; and territorial availability (i.e. it is not an area addressed by existing services in the UK). Whilst the ultimate conclusion to the question of the requirement and utility of this proposed service was a cautious endorsement, the report strongly questioned its overall feasibility (at least in terms of the way that the observatory was envisioned in the original ITT).

The purpose of presenting this work at IPRES is to offer the broader community an opportunity to comment on the costs observatory concept. To facilitate this, further detail will be provided about the conclusions of the Key Perspectives report.

4. THE BLUE RIBBON TASK FORCE ON SUSTAINABLE DIGITAL PRESERVATION AND ACCESS

The third and final strand of work to be included in this paper is an activity that was initiated by the National Science Foundation and the Andrew W. Mellon Foundation in the U.S., but was also supported by a number of other funders including JISC.[13] The purpose of the Task Force was to:

- Conduct an analysis of previous and current models for sustainable digital preservation, and identify current best practices among existing collections, repositories and analogous enterprises
- Develop a set of economically viable recommendations to catalyze the development of reliable strategies for the preservation of digital information
- Provide a research agenda to organize and motivate future work in the specific area of economic sustainability of digital information

The Task Force was convened over a two year period and delivered a significant and influential report in February 2010 that was widely referenced and nominated for the 2010 DPC Preservation award.[14] One of the features of this work that distinguishes it from the preceding projects, but also makes it nicely complementary, is that the focus is not on the 'cost' of digital preservation, but is more to do with the economic factors and strategies that may determine whether it will be possible to sustain digital information in accessible and comprehensible environments for the foreseeable future.

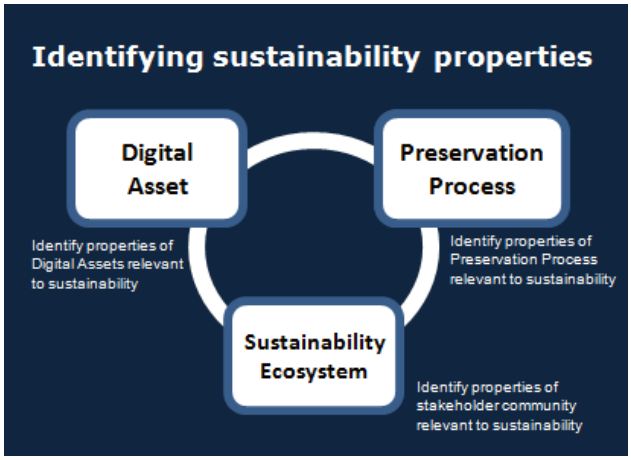
One of the features of the report is that it reframes some of the imperatives of digital preservation into an alternative (economic) language, where the laws of supply and demand, and some more specific language such as describing digital materials as 'depreciable durable assets', and discussing their 'non-rival' nature in terms of presenting a 'free-rider problem' offer a new type of terminology for understanding the challenges associated with managing information.

Whilst this report makes essential reading for a wide range of organisations dealing with a diverse array of data types, it presents the results of two years of detailed and deep thinking into a complex area. After two additional dissemination events in Washington and London, it was apparent that some form of synthesis work was required to present both the conclusions of the BRTF work itself and the subsequent discussions about it.

4.1 The Economic Sustainability Reference Model (ESRM)

In discussion with one of the BRTF panel members (Chris Rusbridge) about the possibility of commissioning some synthesis activity, it became apparent that an alternative idea had been suggested by another panel member (Brian Lavoie - OCLC) to create a different kind of summary of the BRTF conclusions. Building on the approach taken with the OAIS reference model (open archival information system ISO 14721:2003) Lavoie and Rusbridge suggested that a similar (but necessarily different) approach might be taken with the economic framework first outlined in the BRTF report, and that any resulting graphical depiction or conceptual model might not only act as a more concise and immediately descriptive synthesis of the BRTF work, but may also represent a useful and flexible community tool around which an ongoing discussion about economic sustainability might be based. From the outset, it was envisaged that if the framework received community endorsement, then it might provide a foundation for the kind of standards development process that the OAIS reference model underwent.

Figure 3 The current top level components of the ESRM[15]



At the time of writing, a draft version of the ESRM is still in preparation and the only public exposure the idea has had was at a workshop that took place in Tallinn, Estonia in May 2011, in conjunction with the Aligning National Approaches to Digital Preservation Conference.[16] A report from this event is still forthcoming but in summary, the delegates in attendance approved of the approach and endorsed further work to develop the reference model. IPRES represents another opportunity to demonstrate the latest iteration of the model and to elicit feedback about its likely usefulness and relevance to organisations facing

genuine (rather than theoretical) finance-related problems when preserving their digital assets.

5. OVERALL AIMS

Whilst this paper references six distinct (but more or less related) activities for presentation in a fairly short space of time, it should be acknowledged that the three initial activities (LIFE Project, KRDS project, and the BRTF initiative) ought to represent familiar territory to a lot of the IPRES attendees, many of whom will have extensive knowledge of the published literature on preservation. The objective therefore would be to address these activities with cursory descriptions (enough for those not familiar with them to understand their principle purpose) and then to move rapidly onto describing, and where possible evaluating, the new work that has been commissioned to follow up and build on the earlier work.

As noted above, when introducing all three new areas of work, it will be useful to provoke comments, opinions and discussion from the IPRES delegates to feed into the planning and implementation of next phases. It is not yet apparent to JISC whether and how further funding should be directed at any of the three projects and although it seems highly likely that an ongoing investigation into the economics, costs - and perhaps particularly – the benefits of digital preservation, would be appreciated by the broader community, detailed scoping is required. This presentation to IPRES could be an important part of that process.

6. REFERENCES

[1] Lifecycle Information for e-Literature. http://www.life.ac.uk/

[2] Keeping Research Data Safe (KRDS) Project. Phases I and II. http://www.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx

[3] The Blue Ribbon Task Force for Sustainable Digital Preservation and Access: http://brtf.sdsc.edu/

[4] During 2008, the University College London (UCL) e-Prints Repository recorded that the total number of downloads for the two LIFE project outputs available at that time was 1588. This meant that the items were the 9th and 11th most popularly downloaded items in the repository that year. http://discovery.ucl.ac.uk/past_stats/annual-2008.html#top50

[5] Diagram taken from presentation given by Brian Hole (British Library) at the Preservation and Archiving Special Interest Group (PASIG), Madrid, July 2010, available at: http://www.life.ac.uk/3/docs/Hole_pasig_v1.pdf

[6] A recent description of Danish work on the cost of digital Migration is published in the International Journal of Digital Preservation. http://www.ijdc.net/index.php/ijdc/article/viewFile/177/246

[7] LIFE-SHARE project, University of Leeds, http://www.leeds.ac.uk/library/projects/lifeshare/

[8] DCC LIFE Project web page: http://www.dcc.ac.uk/projects/life

[9] Keeping Research Data Safe: http://www.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx

[10] JISC Managing Research Data Programme: http://www.jisc.ac.uk/whatwedo/programmes/mrd.aspx

[11] Beagrie, N., (2011), Keeping Research Data Safe (KRDS) User Guide, p.5: http://www.beagrie.com/KeepingResearchDataSafe_UserGuide_v2.pdf

[12] A JISC Closed ITT (invitation to tender) was circulated on 16th May 2011 to a limited number of consultancies

[13] The Blue Ribbon Task Force on Sustainable Digital Preservation and Access: http://brtf.sdsc.edu/

[14] Digital Preservation Coalition Digital Preservation Award: http://www.dpconline.org/advocacy/awards/581-2010-digital-preservation-award

[15] Lavoie, B., Rusbridge, C., slide shown at the ESDI Roundtable Meeting, 26/05/11, Tallinn.

[16] Economic Sustainability of Digital Information (ESDI) Roundtable Event, 26th May 2011, National Library of Estonia, Tallinn, Estonia, http://www.educopia.org/events/ANADP/ESDI_Roundtable

Long-Term Sustainability of Spatial Data Infrastructures:
A Metadata Framework and Principles of Geo-Archiving

Arif Shaon
Science and Technology Facilities Council, UK
arif.shaon@stfc.ac.uk

Carsten Rönsdorf
Ordnance Survey, UK
Carsten.Roensdorf@ordnancesurvey.co.uk

Urs Gerber
The SWISS Federal Archive, Switzerland
Urs.Gerber@lt.admin.ch

Kai Naumann
Landesarchiv Baden-Württemberg - Staatsarchiv Ludwigsburg, Germany
kai.naumann@la-bw.de

Paul Mason
Ordnance Survey, UK
Paul.Mason@ordnancesurvey.co.uk

Andrew Woolf
The Bureau of Meteorology, Australia
A.Woolf@bom.gov.au

Michael Kirstein
Generaldirektion der Staatlichen Archive Bayerns, Germany
Michael.Kirstein@gda.bayern.de

Marguérite Bos
The SWISS Federal Archive, Switzerland
Marguerite.Bos@bar.admin.ch

Göran Samuelsson
Mid Sweden University, Sweden
goran.samuelsson@miun.se

ABSTRACT

With growing concerns about environmental problems, and an exponential increase in computing capabilities over the last decade, the geospatial community has been producing increasingly voluminous and diverse geographical datasets. Long-term preservation of these geographical data exposed through uniform and interoperable Spatial Data Infrastructures (SDIs) is not typically addressed, but highly important for meeting legislative requirements, the short and long term exploitation of archived data as well as efficiency savings in managing superseded datasets. In this paper, we attempt to set out the path and describe what needs to be done now to future-proof the investment government agencies around the world have made in digital geographic data. We take the INSPIRE SDI as an exemplar to investigate the requirements for ensuring sustained access to geographical data from the perspective of a preservation-aware and INSPIRE-conformant SDI. We also outline a number of principles for the long term retention and preservation of European digital geographic information defined by the EuroSDR Geographic Data Archiving working group. In addition, we present a preservation profile of the ISO 19115 metadata standard to enable recording and exposing important preservation related information about geographical data through large-scale SDIs like INSPIRE.

Keywords

preservation, archive, metadata, INSPIRE, ISO 19115, geographical data.

iPRES2011, Nov. 1–4, 2011, Singapore.

1. INTRODUCTION

Geo-information systems (GIS) have become an indispensable means of storing and analysing geographical data for government, business, and research. In Europe, the National Mapping Agencies (NMAs) and other geographic institutions today experience rising demand for historical geographical data that describe how land, cities and countries have developed over time. Government agencies around the world have invested heavily in this type of geographical data. Unfortunately, high storage costs and difficulties in finding, accessing and delivering older datasets and raster data[1] are making the task of satisfying this demand extremely challenging. Unlike paper maps, digital geographical data without efficient curation and preservation could become unusable within about one decade due to software, hardware or data model obsolescence. Safeguarding today's fundamental geographical data for future generations in order to understand history as well as historic trends needs to be a core objective of the National Mapping Agencies and other data providers.

The European Union INSPIRE Directive[2] aims to address the need for interoperability across the geographical datasets held by its different member states. To facilitate such a high level of interoperability, the directive mandates the adoption of common Implementing Rules (IR) for metadata, data specifications, network services, and data sharing through a pan-European Spatial Data Infrastructure (SDI). While this is an effective way of ensuring interoperability across disparate datasets, it does not guarantee sustainability of those datasets over an indefinite period of time. For instance, INSPIRE does not address ensuring compatibility with future technology or ensuring continued access even after a provider has ceased to exist. To further illustrate, we can consider the specific requirement of INSPIRE for data providers to use the OGC[3] standardised Web

[1] Raster graphical data - http://en.wikipedia.org/wiki/Raster_graphics

[2] INSPIRE Directive - http://inspire.jrc.ec.europa.eu/

[3] Open Geospatial Consortium - http://www.opengeospatial.org

Map Services[4] to expose GIS Maps. Currently, there is no standardised way of defining precisely which data tables, attributes, geometries or raster images are contained within such a service. But each of those components has different properties that will need to be migrated into newer systems or formats at some point in time to ensure continued accessibility and usability.

Geographic information is already at the heart of environmental analysis that informs policy as well as practical implementation. For example, adding preserved digital snapshots of detailed land ownership and use, river and transport networks together with historical environmental measurements such as pollution or water quality over 10 or 20 years, coupled with new analysis techniques not available today will identify correlations and trends that allow better scenario models for the future and also inform environmental policy. As this shows, properly historicised geographic information provides tremendous value for government, economy as well as for individuals. We need historic data to meet economic and legal requirements for government and business, but also for citizens as a means of gaining deeper understanding of their lineage, for example, by tracing back their individual or family history. A large-scale SDI like INSPIRE has a crucial role to play in facilitating the availability of this type of geographical data over the long-term.

In this paper, we investigate the requirements for developing a preservation-aware SDI based on the OAIS reference model [5], an important ISO standard for digital preservation. We also outline a number of principles for the long term retention and preservation of digital geographical information with a view to introduce fundamental concepts of digital geographical data archiving for the public sector information providers in Europe. These principles have been proposed by the EuroSDR Geographic Data Archiving working group[5] – a group of 11 National Mapping Agencies, Archives and Research institutions across Europe collaborating to address the issues of preserving geographical data in Europe. In addition, we present a preservation profile of the ISO 19115 metadata standard[6] that is designed to enable an archive to record preservation-related information about geographical data and make it available to the users through the associated SDI.

2. THE MAIN CHALLENGES OF PRESERVING GEOGRAPHICAL INFORMATION

In general, geographical data inherit the preservation challenges inherent to all digital information [3]. These challenges are further complicated by some of the characteristics of geographical datasets, such as diverse and highly structured data formats, and the need for special domain knowledge for accurate interpretation. Moreover, in the context of SDIs, such as INSPIRE, state-of-the-art service-oriented infrastructures adopt exchange formats (i.e. application schemas) that reflect domain-specific conceptual data models ('feature types') rather than directly reflecting underlying database storage schemas. These application schemas and their relationships (e.g. mapping) with the corresponding datasets would need to be preserved to ensure appropriate accessibility and re-use of those datasets in the future.

On the positive side, it should be possible in principle, to apply existing widely adopted preservation mechanisms and standards, such as the OAIS reference model (Section 4) to the long-term preservation of geospatial data. In fact, a number of European archives [10] are currently adopting or are looking to adopt the OAIS model and other related specifications for the long-term preservation of their geospatial datasets. These organisations would, therefore, significantly benefit from a best-practice implementation profile of the OAIS model for geospatial datasets and an INSPIRE-compliant metadata model for describing and sharing the relevant preservation aspects (Section 5) of such datasets through the INSPIRE SDI – neither of which exist at present.

3. EXISTING ENDEAVOURS

Aside from a handful of initiatives, such as the NGDA[7] project funded by the NDIIPP[8] initiative of the US Library of Congress, the GER[9] project and some exploratory work by the Digital Preservation Coalition (DPC) [3], there have not been many noteworthy endeavours for long-term preservation of geospatial information. Amongst the existing initiatives, the GER project has introduced a new metadata model for describing geospatial information, which is essentially an amalgamation of FGDC[10] (the current US Federal Metadata standard), the ISO 19115 metadata model and a few preservation metadata specifications including the PREMIS Data Dictionary [9]. In general, the GER model is a comprehensive metadata model designed to enable capturing and managing a wide variety of preservation-related information (e.g. accessibility, provenance, distribution etc.) about a geospatial dataset during its entire life-cycle. The metadata-related notions defined in the GER are represented as relational database tables and their corresponding fields, with a view to facilitate the development of new archives for preserving geospatial data as well as improving the capabilities of existing archives [6]. As a result, the GER metadata model is not a true 'profile' of any of the existing metadata standards on which it is based; e.g. it does not follow the rules of profiling specified in Annex C of ISO 19115. From that perspective, it would not be fit for capturing and sharing metadata about geospatial datasets through large-scale SDIs, such as INSPIRE which requires the adoption of ISO 19115-conformant metadata models for describing geospatial data.

The NGDA approach, on the other hand, is specifically intended to address the preservation requirements of the US-based

[4] http://www.opengeospatial.org/standards/wms

[5] EuroSDR Geographic Data Archiving working group - http://bono.hostireland.com/~eurosdr/start/index.php?option=com_content&task=view&id=60&Itemid=88

[6] ISO 19115:2003 Geographic information – Metadata

[7] National Geospatial Digital Archive (NGDA) Project - http://www.digitalpreservation.gov/partners/ngda/ngda.html

[8] National Digital Information Infrastructure and Preservation Program (NDIIPP) - http://www.digitalpreservation.gov/library/

[9] Geospatial Electronic Records (GER) project - http://www.ciesin.columbia.edu/ger/

[10] Federal Geographic Data Committee (FGDC) Metadata Format - http://www.fgdc.gov/metadata

geospatial datasets at archive or repository levels. In particular, this approach includes a comparative assessment of a number of existing metadata standards, including the aforementioned GER and FGDC metadata model with a view to address the metadata capturing and management requirements of a long-term archive of geographical information [1]. However, such archive-specific technical solutions may not directly benefit large-scale SDIs more generally (including INSPIRE), where the main focus is on the provision of uniform accessibility of geospatial datasets, not specific techniques for preserving such datasets. Further, an SDI typically consist of many different data providers with different organisational remits and constraints – so, it would be impractical for an SDI to impose the adoption of a 'one-size-fits-all' preservation approach on all the data providers involved. Nevertheless, the NGDA approach could serve as guidelines for implementing geospatial preservation archives in Europe, mainly for the exploratory work done on various general aspects (e.g. data format, metadata mapping etc.) of long-term preservation of geospatial data.

Aside from the aforementioned endeavours, the European Space Agency (ESA) has recently established a major preservation initiative, the ESA Long-Term Digital Preservation (LTDP)[11] programme, with a view to formulate a coordinated and coherent approach to the long-term preservation of the EO space data archives across its member states. Although this ESA LTDP initiative primarily focuses on the preservation of Earth Observation (EO) space data, the end result of this initiative should also be applicable to other types of geographical data, and to INSPIRE. The work presented in this paper should be of considerable relevance to this ESA initiative, since ESA adopts ISO 19115 for collection-level discovery[12].

4. THE OAIS REFERENCE MODEL

The Reference Model for an Open Archival Information System (OAIS) is a very important ISO standard (ISO 14721:2003) for addressing the issues associated with the long-term preservation of digitally encoded information [5]. The OAIS describes a number of conceptual models in order to aid formulation of a suitable preservation strategy for digital objects. Of particular importance, among the OAIS models, is the Information Model that broadly describes the metadata requirements associated with retaining a digital object over the long-term (Figure 1). We consider the different components of the OAIS information model from the perspective of long-term preservation of geospatial datasets.

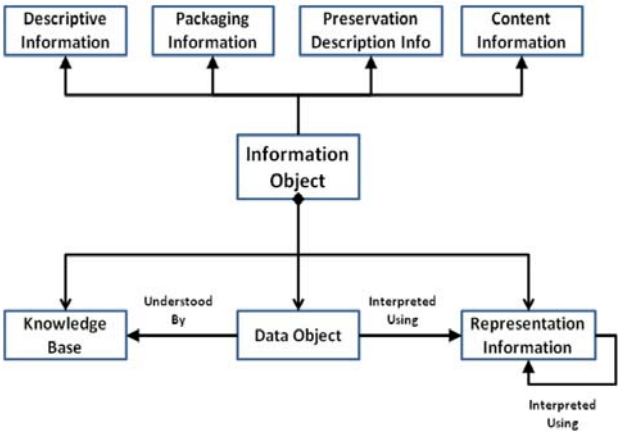


Figure 1: A Partial View of the OAIS Information Model [5]

4.1 Content Information

This is the set of information that needs to be preserved over the long-term. In the case of spatial datasets, it should be the 'original' version of a dataset rather than a domain specific representation of that dataset. For example, in the INSPIRE SDI, where geospatial datasets are mapped on to 'application schema' to represent particular facets of phenomena on the earth as 'geographic features' (e.g. a pan-European road transport network), the source dataset rather than its 'mapped view' should form the 'Content Information'.

4.2 Preservation Description Information (PDI)

This type of information is needed to efficiently manage and preserve a digital object over an indefinite period of time. This includes various information about the life-cycle of a dataset, such as its provenance and versioning history, as well as reference and annotation-related information.

4.3 Representation Information (RI)

This is a component of the Content Information that is required to accurately render a preserved digital object on a future technological platform. This encompasses all levels of abstraction and refers to both the structural and semantic composition, such as recreating the original appearance of the digital object, or analysing it for a concordance [5]. The use of RI can be recursive, especially in cases where meaningful interpretation of one RI element requires further RI (Figure 1). The RI for a dataset may include information about its technical dependencies, such as software required to access the dataset, compatible operating platform and so on.

With respect to an SDI, RI refers to the ability to continue to be able to interpret the semantics of a digital dataset, i.e. how the digital objects relate to a conceptual model of some universe of discourse (ISO 19101:2002 - Geographic information -- Reference model). For instance, a transport network dataset stored in a geo-database or a Shapefile[13] will be meaningless unless the tables or digital objects can be interpreted as 'road features' defined in a relevant conceptual model.

[11] European Space Science (ESA) Long-Term Digital Preservation (LTDP) Programme - http://earth.esa.int/gscb/ltdp/. (See also: http://www.digitalpreservationeurope.eu/publications/briefs/dp_for_longterm_environmental_monitoring.pdf)

[12] ESA HMA Standards - http://earth.esa.int/gscb/HMAstandards.html

[13] Shapefile - http://en.wikipedia.org/wiki/Shapefile

4.4 Packaging Information

This type of information is used to bind a data object and its associated metadata (such as PDI and Descriptive Information) into an identifiable unit or package for preservation. For example, if a data object is compressed before being ingested into an archive, the packaging information for that dataset would include information about the underlying structure of its compressed form.

4.5 Descriptive Information

The information needed to facilitate efficient discovery and accessibility of a preserved data object, typically through search and retrieval facility provided by the long-term preservation archive. Descriptive information about a data object may be derived from its PDI and other metadata. For a spatial dataset that is exposed as a 'feature type' through for example, an OGC standardised Web Feature Service (WFS)[14], the descriptive information could include the information (e.g. keywords, abstract) about that 'feature type' provided in the 'GetCapabilities' document of the WFS.

4.6 Designated Community/Knowledge Base

This encompasses all identified potential consumers (e.g. human, software application etc.) to whom the preserved data object is beneficial in terms of its accurate interpretation and proper utilisation. The level of recursion for a particular element of representation information (RI) about a data object is likely to depend on the level of knowledge that the designated community has about that element. For example, if the designate community has considerable understanding of the OGC Web Feature Service, then the representation information of a dataset that is exposed through WFS as 'feature types' could just include the service name – 'OGC Web Feature Service'. Conversely, if the designated community has no understanding of WFS, the representation information of such dataset would have to include detailed implementation and use specification of the OGC WFS among other related information.

A generic viewpoint assumption in an SDI for long-term preservation would define the user community of the SDI as the OAIS 'designated community', with the semantics of harmonised conceptual models that enable domain-specific representation (e.g. 'feature types') of a spatial dataset within the SDI constituting the OAIS 'knowledge base'.

5. A PRESERVATION-AWARE SPATIAL DATA INFRASTRUCTURE

We have analysed the INSPIRE architecture in the context of the OAIS reference model with a view to determining the requirements for a preservation-aware SDI. Functionally, INSPIRE consists of the following components (Figure 2):

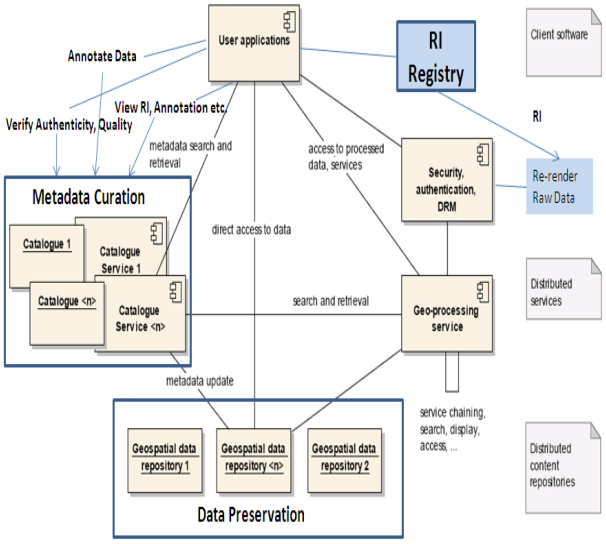


Figure 2: A Preservation-aware SDI

- **Geospatial Data repositories** made available and maintained by different member states and other approved data providers.
- **Metadata catalogues** containing metadata - additional information about the data held in the repositories, typically provided by the data provider(s) - based on the ISO 19115 metadata model to enable efficient discovery of the data exposed through the repositories.
- **Geo-processing Web services** to enable accessing, analysis and processing of the data discovered using the metadata catalogues; includes view and download services.
- **User applications**, i.e. client software to enable users to search the metadata catalogues in order to locate datasets for further processing and/or analysis using the geo-processing services as required.

An analysis of the applicability of the OAIS reference model to the INSPIRE SDI identifies the following three core requirements for ensuring sustained accessibility and usability of the data exposed through such SDIs.

5.1 Long-term preservation of geospatial data repositories

An effective and coherent approach is required to preserve the individual data repositories made available through the SDI over the long-term (Figure 2 – "Data Preservation" box). This needs to address various complex issues, such as compatibility of data with future repository technology and ensuring its continued access even after its provider has ceased to exist. While this aspect is provider-specific, and dependent on the adoption of suitable preservation policies and strategies, it should be possible for the repository owners to identify, define and adopt a set of common fundamental concepts or principles of archiving geographical data over the long-term. INSPIRE can play an important role in defining and promoting such preservation concepts and principles, or at the very least, creating an

[14] OGC Web Feature Service - http://www.opengeospatial.org/standards/wfs

awareness of the importance of long-term preservation of geographical data among the data providers.

5.2 Preservation-aware Metadata Model

The ISO 19115 metadata model adopted in the INSPIRE SDI is comprehensive enough for capturing enough of the context surrounding the data (for example, data quality, maintenance, use/processing) to enable its effective discovery. However, the metadata elements defined in ISO 19115 do not capture other important preservation-related metadata specified in the OAIS Reference model, such as PDI and RI (Section 4). For example, the ISO 19115 model does not address the mappings between a source geospatial data set and its canonical representation, which typically describes particular facets of phenomena on the earth as 'geographic features'. Such 'feature-based' representation of a geospatial dataset is usually described by an appropriate 'application schema' and exposed by the INSPIRE SDI. This type of information is a significant aspect of a geospatial dataset's RI, without which accurate interpretation and re-use of the dataset on a future technological platform may not be possible.

Therefore, a preservation-aware SDI would require a preservation-focused metadata model that would help capture accurate and sufficient description of all aspects (including the aforementioned preservation-related aspects) of a geospatial dataset as well as well as being flexible for addition of future requirements. However, as RI of a dataset could be highly complex and detailed (depending on the requirement of the designated community), it may be sufficient for a preservation metadata model for a SDI to include only an overview of the RI associated with a dataset. Access to the complete set of RI could be provided through a RI repository or registry (Figure 2), if supported by the data provider. There are other benefits in adopting such an approach that are discussed in Section 7.1.

5.3 Long-term curation of metadata catalogues

The metadata catalogues (Figure 2 – "Metadata Curation" box) are instrumental in facilitating discovery of the datasets held in the repositories by enabling searching of the metadata that describe those datasets. However, without curation - proper management, quality assurance and preservation - the metadata, too, may become unusable over time (Figure 2 – "Metadata Curation" box). For example, it may become out of step with the data that it describes. Therefore, it is also crucial to apply effective long-term curation measures to the metadata catalogues within an SDI [8].

6. PRINCIPLES OF ARCHIVING GEOGRAPHICAL DATA

As mentioned before, the data providers of large-scale SDIs, such as INSPIRE should benefit from a set of common and practical principles applicable to the task preserving geographical data over the long term.

In recognition of the importance of long-term archiving of geographical data in Europe, a number of National Mapping Agencies, archives and research councils across Europe formed the EuroSDR Geographic Data Archiving working group in 2010. Since its inception, the group has been working together to identify, articulate and address the challenges faced by European data providers for preserving their geographical data. As an outcome of this exercise, the group has recently defined and agreed upon a set of common and practical fundamental concepts and principles of archiving geographical data.

Here, we outline a selected few of these principles as agreed by some of the important European National Mapping Agencies and archives who expose their geographical data through SDIs like INSPIRE.

The order of the principles follows the lifecycle of data from creation to maintenance, archival, preservation to accessing archived data. Notably, more generic and comprehensive conceptualisations of the lifecycle of an archive already exist. For example, the Draft DCC Curation Lifecyle Model has been designed to facilitate a lifecycle approach to the management of digital materials in an archive, and to enable their successful curation and preservation from initial selection for reuse and long-term preservation [11]. The principles presented here are the outcomes of a preliminary exploration of the applicability of these existing models to Geographical archives.

Suggested action points in the principles are indicated by this symbol: ►.

Principle 1: *Archiving of digital geographic information begins at the point of data creation, rather than at the point of withdrawal from active systems.*

Today archiving is often seen as an afterthought, though the long term value of a dataset can often be appraised at the outset. If this is done, archival requirements are clear from the start and can be acted upon.

► Define whether long term preservation is desired or necessary, determine and document the retention period. This can be changed at a later date if requirements change but will clarify archival needs from the outset. It should also be done for all existing datasets.

Principle 2: *Establishment and agreement of a common preservation planning process and a set of common preservation objectives between data producers and archives is the backbone for any archiving business case.*

► An archive should look across borders and beyond its domain, and consult other experts to formulate an efficient preservation strategy. Using a common vocabulary and reference model (such as the OAIS model) will improve clarity and understanding. One of the key goals of a long term archiving/preservation strategy is risk mitigation against loss and corruption.

► The preservation objectives of an archive should be defined and articulated in its archival policy. The policy should cater for the requirements of both data providers and future users (the so-called designated community).

► A good governance regime is needed to be established to ensure that the policy is implemented in the foreseeable future.

Principle 3: *Be selective and decide what to archive and what to lose.*

Archiving is an economic issue, as well as a technical challenge. Long term benefits are likely to be intangible, so it is advisable to concentrate on short and medium term benefits. Long term archiving may prove to be less challenging if the medium term actions are considered, prepared and undertaken well. The

survival rate for data might be better if less material is archived well, than a vast amount of material being archived poorly.

► An archive should define for each dataset, product or feature group, the required retention period. It should also preserve the documentation that explains what it has chosen to lose and why. This means that it needs to be explained why which aspects of a dataset are important in the shorter and longer term (collection policy).

Figure 3: Geo-archiving Lifecycle

Principle 4: *Consider archiving timeframes of 1, 10, 100 years*

1 year, operational archives focus on short term needs, proprietary formats and specialist solutions may be appropriate.

10 years, a strategic, internal business archive, the focus should be on reusability and access of data. This builds a bridge between shorter term data provider's needs and archivists' needs.

100 or even 1000 years, long-term archive aimed at preservation. Focus on robustness against data loss and corruption, ability to curate and migrate. Data preferably held in flat files, open format.

► Planning should be made to shift data between these archives which may be based on different technical solutions. Access to the 100 year archive can be through a replicated data in a 10 year archive.

Principle 5: *The output of the planning process should also be preserved over the long-term to accommodate future preservation requirements.*

► The documents describing the archival planning process and policy need to be linked to the geographic data in order to provide the context for decision made at the time at or before ingestion of data into an archive.

Principle 6: *Archiving is not backup.*

► It is necessary to backup an archive on at least two uncorrelated storage systems. One backup system should be at a remote and secure site.

Principle 7: *Geographical data should be preserved in a way that non geo-specialists can handle it.*

The likelihood that data survives and can be accessed will be higher if data is structured in a way that archivist are familiar with from other, non-geospatial mainstream content.

► Document migrations, format, and structure so it can be understood by archivists and curators.

► Document the motivation behind applying certain preservation action (e.g. migration) to the data. This type of information forms the preservation history of a dataset and may assist future archivists in understanding and determining the updated preservation requirements for that dataset.

► Also archive data specifications, definitions of coordinate systems and anecdotal material that will help to interpret and understand the data at a later point in time.

Principle 8: *Ensure effective management and quality assurance of the metadata associated with your data.*

► Define the types of metadata needed to enable efficient discovery, accurate rendering, understanding and re-use (e.g. significant properties), and effective preservation of your data over the long-term

► Use appropriate, widely-adopted metadata standards and formats (e.g. ISO 19115, Dublin Core[15], ISO 23081[16])

►Metadata stored in the archive should be both syntactically and semantically valid. For example, an XML-based metadata record can be validated the corresponding XML schema to ensure structure validity. Semantic validation is more complex, and may involve the use of controlled vocabulary defined by the archive, preferably through collaboration with the user community.

► Apply appropriate and efficient versioning mechanism to manage changes made to the metadata in the archive over time.

► Consider enabling the users to annotate the metadata in the archive to facilitate adding value to the metadata.

► Define a set of broad and high-level principles that form the guiding framework within which the metadata curation (management) can operate. The metadata curation policy would normally be a subsidiary policy of the archival data preservation policy statements and should have reference to the rules concerning legal and other related issues regarding the use and preservation of data and metadata, as governed by the data policy statements.

7. A PRESERVATION PROFILE OF THE ISO 19115 METADATA MODEL

As identified in the analysis of the INSPIRE SDI above (Section 5), the ISO 19115 metadata model is not sufficient for capturing and providing the users with the information needed to enable accurate interpretation of geospatial data in the future. To address this issue, we have developed a preservation profile of ISO 19115 based on the metadata requirements specified in the OAIS reference model and the PREMIS data dictionary[17]. The

[15] Dublin Core Metadata Elements Set - http://dublincore.org/documents/dces/

[16] ISO 23081: Records Management Processes - Metadata for Records

[17] A framework for defining and describing a set of core preservation metadata (based on the OAIS reference model)

rationale of this profile is to enable recording preservation-related information about a geospatial dataset, while retaining the ability of the core ISO 19115 model to capture descriptive and contextual metadata about that dataset. The preservation profile incorporates the following key preservation concepts into the core ISO 19115 model as shown in Figure 4 below.

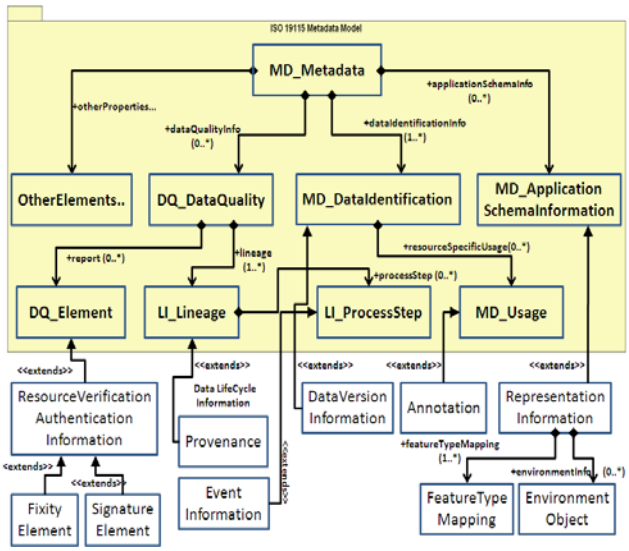


Figure 4: A preservation profile of ISO 19115 Metadata Model

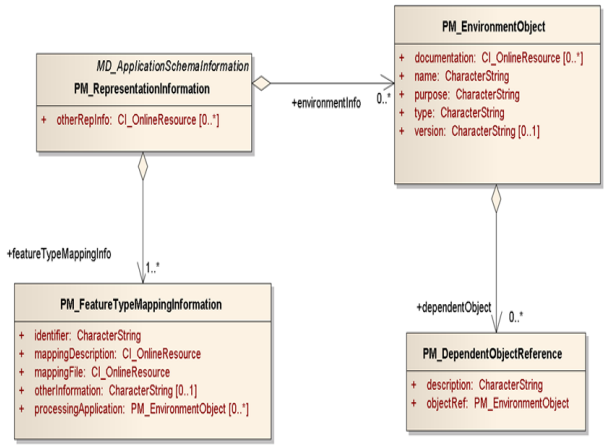


Figure 5: Representation Information elements of the preservation profile of ISO 19115 Metadata Model

7.1 Representation Information

The OAIS reference model defines the Representation Information (RI) about a digital object as the information required to enable access to preserved digital objects in a meaningful way [5]. In ISO 19115, the only notable RI related information defined is the information about the application schema(s) (i.e. the ***MD_ApplicationSchemaInformation*** class – Figure 3) used to create a particular feature view of a source geospatial dataset. The preservation profile extends this concept to incorporate information about the mappings between the source data and application schema along with the applications/software/services required to effectively apply the mappings (Figure 4).

In particular, as illustrated in Figure 5, the preservation profile defines the ***PM_FeatureTypeMappingInfo*** class to record information about the mapping(s) between a source dataset and its canonical 'feature-based' representation. The preservation profile also defines additional elements (***otherRepInfo*** and ***environmentInfo*** properties of ***PM_RepresentationInformation*** class – Figure 4) to enable capturing other data specific RI (e.g. data formats, storage media), in the form of web-accessible resources (through HTTP URLs). It is envisaged that detailed RI about a geospatial dataset may not directly benefit its typical users, as they are likely to rely on the current data provider or preservation body to make the data available to them, generally through web services, which apply the aforementioned mappings.

Nevertheless, this approach provides the users with the option to access the RI (made available on the web through e.g. a RI registry by the data provider/preservation body) about a dataset, which, if necessary, could be used to reconstruct and re-use that dataset on a future technological platform (Figure 2). From an archivist's perspective, it is an important mechanism for providing access to the data in a consistent manner into the future. As well, it provides flexibility in terms of the metadata model/format used to capture data-specific RI without being constrained by the ISO 19115 model.

7.2 Data life cycle information

Detailed information about changes (e.g. change of ownership or archive) and events occurring during the life-cycle of a dataset is essential for verifying the provenance of a dataset as well as the reliability of its preservation in the future. In addition, this type of information could contain a detailed history of every preservation measure (e.g. migration) applied to a dataset during its lifecycle, in order to assist its future curators in understanding and determining the updated preservation requirements for that dataset. For instance, a provider may choose to migrate an existing road transport dataset into a new database schema more closely reflecting an INSPIRE application schema (a process sometimes known as 'Extraction-Transformation-Load', or ETL); it is important to document this schema transformation for preservation purposes. Similarly for quality assurance purposes it is important to be able to verify the history of ownership of a dataset.

With this in mind, the preservation profile extends the ***LI_Lineage*** and ***LI_ProcessStep*** elements (Figure 4) defined in the ISO 19115 model to capture detailed information about the lifecycle of a dataset. The dataset lifecycle information in the preservation profile is divided into two main categories: **Dataset Provenance Information**, (i.e. change of ownership and/or preservation body) and **Dataset Event Information** (i.e. all major events, including preservation-related ones, such as major platform change and preservation certification process that have affected the data during its life cycle - useful for audit trailing and quality checking purposes).

that would be required to facilitate a long-term data preservation process in a digital archive [9].

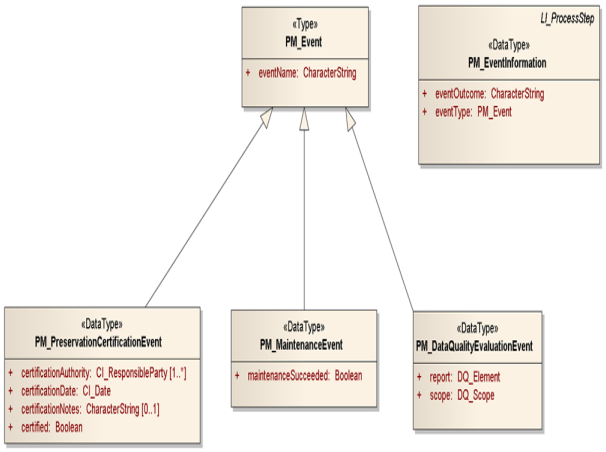


Figure 6: Dataset Event information elements of the ISO 19115 Preservation Profile

Important among these elements is the ***PM_PreservationCertificationEvent*** (a specialised ***PM_Event*** class shown in Figure 6) defined to provide information about any certification examination(s) conducted, to ensure adequacy of the preservation measure(s) applied to a dataset. This should provide the users with some level of confidence in the preservation method(s) applied to, and consequently, in the longevity of the data of their interest. In the OAIS, this type of information is referred to as 'Preservation Descriptive Information' (See Section 4.2).

7.3 Data Authenticity Verification Information

The ISO 19115 model adopts a number of data quality related concepts (e.g. ***DQ_Elements*** – Figure 4) from the ISO 19113[18] and 19114[19] standards (for representing the quality principles and evaluation procedures associated with geographic information) in order to provide detailed description of the quality assurance measures applied to a dataset. The preservation profile adds to this the ability to verify unauthorised modifications to a dataset by recording its fixity information, such as a checksum and digital signature. This may be important, for instance, where major asset management or security programmes depend on the accuracy of information in a dataset, and it is important to be sure that data has not been altered.

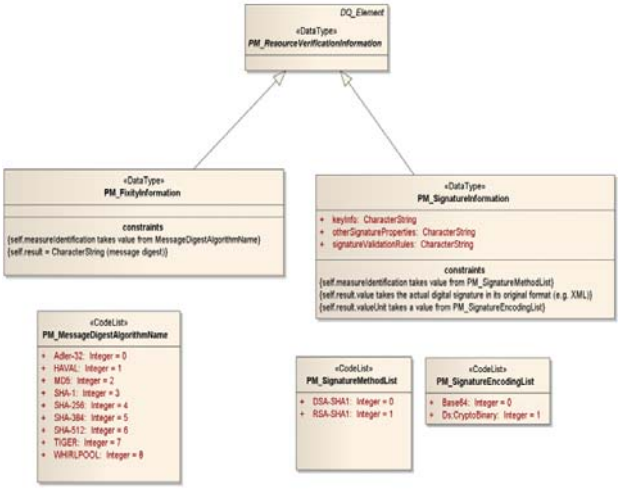


Figure 7: Resource Authenticity Verification information elements of the ISO 19115 Preservation Profile

As illustrated in Figure 7, the preservation profile defines the ***PM_ResourceVerificationInformation*** class as a specialised ***DQ_Element*** class (of ISO 19115:2003 core). It is intended to record fixity information (***PM_FixityInformation*** class), such as a checksum and digital signature (***PM_SignatureInformation*** class) about a dataset to enable verification of unauthorised alterations made to that dataset.

In the context of the OAIS information model, this type of information is categorised as the 'Preservation Descriptive Information' associated with a dataset.

7.4 Annotation

Annotation in the digital world has long been recognised as an effective means of adding value to digital information. It can, in effect, help establish collaborative links between data providers, data users and a preservation body. Thus, annotation has the potential to facilitate enhanced efficiency of a preservation process, and thereby improve the quality of both data and metadata. However, annotation without the intended context may become meaningless. For example, an annotation may be used to label particular map features with descriptive text, which may contain values of some attributes associated those features [7]. These attribute values alone, i.e. without the correct association with the corresponding map features (the annotation context) would be meaningless. For more complex and dynamic geographical datasets, it may be useful for users to be able to annotate specific features or attributes for collaborative analysis or interpretation, for instance in an emergency response scenario. While not directly related to preservation, it is not difficult to appreciate the long-term value of such information, e.g. during post-disaster audit of response capability.

[18] ISO 19113:2002 - Geographic information -- Quality principles

[19] ISO 19114:2003 - Geographic information -- Quality evaluation procedures

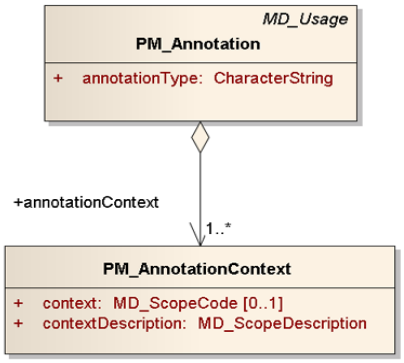


Figure 8: Annotation elements of the ISO 19115 Preservation Profile

Therefore, the preservation profile defines as extensions to the ***MD_Usage*** elements of the core ISO 19115 (Figure 8) a number of suitably structured elements to capture detailed annotation related information (***PM_Annotation*** class) with traceability to the data context (***PM_AnnotationContext*** class**)** to which the annotation refers.

7.5 A Test Case

We tested the preservation profile of the ISO 19115 by recording preservation metadata about some weather observation datasets exposed by an OGC-compliant Web Feature Service (WFS). This WFS is built on the 'Complex Datastore' version of GeoServer[20], which enables representation of data from a relational database in a GML[21]-based application schema (e.g. Climate Science Modelling Language, CSML[22]) defined independently of the underlying database structure. This special edition of GeoServer was a research endeavour by SeeGrid[23] with contribution from the GeoServer community.

Considering the aforementioned special capability of the WFS, the dataset exposed by it provided ideal examples of 'feature-based' representations of source spatial datasets. Therefore, we used the preservation profile of ISO 19115 to record a number of useful Representation Information (RI) about some of the datasets served up by the WFS. This RI captured included the mappings used to generate a "feature-based" canonical representation of a dataset as well as other metadata. The following XML snippet provides an example of such an RI:

```
<geop:PM_RepresentationInformation>
        <geop:featureTypeMappingInfo>
                <geop:PM_FeatureTypeMappingInformation>
                        <gco:identifier>
                                <gco:CharacterString>
                                mapping1
                                </gco:CharacterString>
                        </gco:identifier>
                <geop:mappingDescription>
                                <gmd:CI_OnlineResource>
                                        <gmd:linkage>
        <gmd:URL>http://www.stfc.ac.uk/geopres/mappings/dataset1/descri
ption.html</gmd:URL></gmd:linkage>
                </gmd:CI_OnlineResource>
                        </geop:mappingDescription>
                        <geop:mappingFile>
                                <gmd:CI_OnlineResource>
                        <gmd:linkage>
        <gmd:URL>http://www.stfc.ac.uk/geopres/mappings/dataset1/
mapping.xml </gmd:URL></gmd:linkage>
                </gmd:CI_OnlineResource>
</geop:mappingFile>
<geop:processingApplication>
        <geop:PM_EnvironmentObject>
                <geop:documentation>
                        <gmd:CI_OnlineResource>
                                <gmd:linkage>
        <gmd:URL>http://www.stfc.ac.uk/geopres/mappings/dataset1/a
pplication.html </gmd:URL>
        </gmd:linkage></gmd:CI_OnlineResource></geop:documentation>
        <geop:name><gco:CharacterString>GeoServer WFS
        </gco:CharacterString></geop:name>
        <geop:purpose>
        <gco:CharacterString>produces representation of STFC sample
weather observation datasets in Climate Science Modelling Language - a GML-
based application schema</gco:CharacterString></geop:purpose>
<geop:type><gco:CharacterString>Software</gco:CharacterString></geop:type>
<geop:version> <gco:CharacterString>Complex
Datastore</gco:CharacterString></geop:version>
</geop:PM_EnvironmentObject>
</geop:processingApplication>
</geop:PM_FeatureTypeMappingInformation>
</geop:featureTypeMappingInfo>
</geop:PM_RepresentationInformation>
```

Listing 1: an example of Representation Information recorded using the ISO 19115 Preservation Profile

Of particular note in the above XML snippet is the 'CI_OnlineResource' related metadata elements, such as 'mappingFile' and 'processingApplication'. These elements are defined to record references to web-based resources providing more comprehensive (and possibly complex) information about the aspects of the data that they represent. In the above XML snippet, the 'processingApplication' element points to a web-based document providing detailed information about the GeoServer WFS, such as the input parameters and computer platform required to apply the mappings (described by the 'mappingDescription' and 'mappingFile' elements) to the corresponding dataset. These web-based resources could be encoded in any format chosen by the preservation body concerned. Thus, the preservation profile of ISO 19115 provides flexibility in terms of the metadata model/format used to capture data-specific RI without being constrained by the ISO 19115 model while ensuring the accessibility of such information in a uniform and coherent manner.

[20] GeoServer, an open source Java-based web server that provides a suitable means of promoting and publishing Geospatial information on the web using various OGC standards - http://geoserver.org/display/GEOS/Welcome [Accessed 1 February 2011]

[21] Geography Markup Language is an XML grammar written in XML Schema for the description of application schemas as well as the transport and storage of geographic information - http://www.opengeospatial.org/standards/gml [Accessed 1 February 2011]

[22] http://ndg.nerc.ac.uk/csml/ [Accessed 1 February 2011]

[23] https://www.seegrid.csiro.au [Accessed 1 February 2011]

8. CONCLUSIONS AND FUTURE DIRECTION

Long-term preservation of geographic data exposed through uniform and interoperable SDIs is not currently addressed in the INSPIRE Directive but is highly important for applications that require continued access to both current and historical data e.g. for monitoring climate change. The main drivers for archiving digital geographic information are meeting legislative requirements, the short and long term exploitation of archived data as well as efficiency savings in managing superseded datasets. This paper has attempted to set out the path and describes what needs to be done now to future-proof the investment government agencies around the world have made in digital Geographic Data.

In this paper, we have investigated the requirements for ensuring sustained access to geographical data from the perspective of a preservation-aware and INSPIRE-conformant SDI. We have also outlined a number of principles for the long term retention and preservation of digital geographic information defined by the EuroSDR Geographic Data Archiving working group with a view to introduce fundamental concepts of digital geographic data archiving for the public sector information providers in Europe. In addition, we have presented a preservation profile of the ISO 19115 metadata standard to enable an archive to record preservation-related information about geo-data and make it available to the users through the associated SDI.

Future work in this area would need to focus on the implementation of efficient and interoperable preservation solutions for the data repositories made available through the SDI. To that end, the EuroSDR group aims to define a reference implementation profile of the OAIS reference model for geographical data based on the practical preservation related use-cases extracted from the participating archives and NMAs. A key consideration of this work will be to consider risks and issues for curation and preservation of geographic data throughout the archival phase of its lifecycle [11]. The group will also work towards refining the preservation principles presented in this paper through broader engagement with the NMAs and archives as well as other preservation-related endeavours in Europe.

9. ACKNOWLEDGMENTS

The work presented in this paper was funded in part by the e-Science centre, STFC.

10. REFERENCES

[1] Hoebelheinrich, N. and Banning, J., 2008. An Investigation into Metadata for Long-term Geospatial Formats. *NGDA* Report, (2008) URL= http://www.digitalpreservation.gov/news/events/ndiipp_meetings/ndiipp08/docs/session7_hoebelheinrich_paper.doc [Accessed 4 February 2011]

[2] Janée, G., Mathena, J. and Frew, J., 2008. A Data Model and Architecture for Long-term Preservation. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL),*134–144. (2008) DOI:http://dx.doi.org/10.1145/1378889.1378912

[3] McGarva, G., Morris, S. and Janée, G., 2008. Preserving Geospatial Data, *Technology Watch Report, Digital Preservation Coalition (DPC), DPC Technology Watch Series Report* 09-01. (2008) URL= http://www.dpconline.org/technology-watch-reports/download-document/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-gred-greg-janee.html [Accessed 4 February 2011]

[4] Morris, S. P., 2006. Geospatial Web services and geoarchiving: New opportunities and challenges in geographic information services. *Library Trends*, 55, pp. 285-303. (2006) URL= http://www.lib.ncsu.edu/ncgdap/documents/MorrisLibraryTrendsFall2006.pdf [Accessed 4 February 2011]

[5] 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)* URL= http://public.ccsds.org/publications/archive/650x0b1.pdf [Accessed 4 February 2011]

[6] GER, 2005. Data Model for Managing and Preserving Geospatial Electronic Records Version 1.00 [, *Center for International Earth Science Information Network (CIESIN) Columbia University*. (2005) URL= http://www.ciesin.columbia.edu/ger/DataModelV1_20050620.pdf [Accessed 4 February 2011]

[7] Bose, R and Reitsma, F., 2005. Advancing Geospatial Data Curation, Conference on Ensuring Long-term Preservation and Adding Value to Scientific and Technical Data, *online papers archived by the Institute of Geography, School of Geosciences, University of Edinburgh*. (2005) URL= http://www.era.lib.ed.ac.uk/bitstream/1842/1074/1/freitsma003.pdf [Accessed 4 February 2011]

[8] Shaon, A and Woolf, A. 2008. An OAIS Based Approach to Effective Long-term Digital Metadata Curation, *Computer and Information Science*, 1(2), 2-12. (2008) URL= http://www.ccsenet.org/journal/index.php/cis/article/download/90/79

[9] PREMIS, 2008. PREMIS Data Dictionary for Preservation Metadata, version 2.0 , *PREMIS Editorial Committee*, (2008), URL= http://www.loc.gov/standards/premis/v2/premis-2-0.pdf [Accessed 4 February 2011]

[10] Bos, M, Gollin, H, Gerber, U., Leuthold, J. and Meyer, U. 2010. Archiving of geodata, A joint preliminary study by swisstopo and the Swiss Federal Archive, *SWISS Archive,*(2010),URL= http://www.swisstopo.admin.ch/internet/swisstopo/en/home/topics/geodata/geoarchive.parsysrelated1.59693.downloadList.93958.DownloadFile.tmp/preliminarystudyarchivingofgeodata.pdf [Accessed 11 September 2011]

[11] Higgins, S. 2008: The DCC Curation Lifecycle Model, *The International Journal of Digital Curation.* Issue 1, Volume 2 (June. 2008), URL= http://www.ijdc.net/index.php/ijdc/article/viewFile/69/48 [Accessed 29 September 2011]

Short Term Preservation for Software Industry

Daniel Draws, Sven Euteneuer, Daniel Simon, Frank Simon
SQS Research
Stollwerckstraße 11
D-51149 Cologne, Germany
+49 2203 9154 0

{daniel.draws,sven.euteneuer,daniel.simon,frank.simon}@sqs.com

ABSTRACT

In today's literature digital preservation and its concepts are usually connoted with long term views on the lifecycle of IT systems and software. In addition to that long term view we believe that concepts available for digital preservation are also useful in short term views where the life span of systems and software is limited to a significantly shorter timeline. In this paper we discuss three different real-world use cases that benefit from DP concepts on a short term basis.

Keywords

Software Escrow, Short Term Digital Preservation, Quality Risk Management

1. OVERVIEW

In most literature digital preservation (DP) is associated with a very long term view on systems: According to the Digital Preservation Coalition it is defined as the "series of managed activities necessary to ensure continued access to digital materials for as long as necessary" [2]. This paper utilises the general ideas of digital preservation for shorter term use cases, such as software escrow and due diligence. The aim is to demonstrate that DP's capabilities are important not only in large scale, long term projects but to extend DP to a much wider range of project types and a broad customer base making use of outsourced IT development activities and delivery of IT services. The demonstration of a commercial use case for DP is another goal of the paper.

1.1 Background Scenario "IT Outsourcing"

The basic scenario for our short-term digital preservation addresses the well-established concept of outsourcing that aims to "subcontract responsibility for all or part of an IT function to a third-party service provider that managed and operates the work" [8]. Today, over 7% of all IT-budgets are spent towards outsourcing contracts and this ratio will – accordingly to analyses by Gartner – increase dramatically to 25% for 2020. Interestingly enough the current hype of cloud computing is one specific type of outsourcing and will account for 70% of the overall outsourcing budgets in 2020.

The fundamental concept for all outsourcing contracts is to delegate responsibility (and risks) to a third party. The advantages of doing so are obvious:

- From a client perspective outsourcing enables focusing oneself on one's core competencies in business. For instance, for an insurance company software development and test or IT operations are not core competencies and therefore instead of retaining complete IT testing or IT operations departments they could be subjected to outsourcing them to a specialised third party.
- Outsourcing providers usually have specialised in their fields and can leverage cross-customer synergies and provide more expertise. Expectations are that providers will be able to deliver a service in a more efficient and effective way at a higher level of quality.
- Since outsourcing needs some level of standardisation with regards to definition of services and interfaces between the involved organisations it usually fosters more advanced payment models, i.e. paying per transaction, per value added or per outsourced process step. This facilitates commercial planning processes and budgeting.

On the other hand, delegating responsibilities introduces new risks as an undesired side effect: purchasers become dependent on external providers. These risks need to be managed pro-actively: What happens if an outsourcing provider goes bankrupt or if it is acquired by another company and the new company discontinues this service? What if prices are increased without any justification? Usually, the purchaser only has a black box view onto the service provider with a clear focus solely on the "what to deliver". The service provider is the only stakeholder knowing "how to deliver" and access to this knowledge is at risk if the service provider terminates the contract. Without the specific knowledge it is difficult to keep the service alive, e.g. by handing over the service delivery to another service provider – or maybe by insourcing it again. The two standard risk mitigation approaches are "software escrow" for the case of the provider going out of business and "due diligence" for the case of insourcing the service at a fair price.

The aforementioned risks and their respective mitigation approaches provide the background to DP in a short term perspective: If DP allows for "ensuring continued access" (to IT systems and services) it can be applied to mitigate risks of outsourcing contracts by limiting the impact of third party dependencies. If an outsourcing contract is complemented by a properly set-up DP initiative, the impacts of providers going bankrupt are limited since the DP activities ensure required knowledge is preserved and ready to be transferred to a different party. The challenge shifts towards assuring the stored information is complete and up to date rather than to preserving for a long period of time.

iPRES2011, Nov. 1–4, 2011, Singapore.

1.2 Overview of the document

This background scenario laid out in the previous section is utilised for the structure of the rest of the document: In Section 2 an established mitigation concept called *software escrow* covering risks associated with providers is presented, and the limitations of the current approach in practise are explained – being the reason for make use of DP. In Section 3 the specific DP concepts required to meet these challenges are revisited in the context of software escrow. In Section 4 an improved software escrow service utilising short-term digital preservation is laid out. In Section 6, several real-world use cases are discussed in the light of this improved concept. The paper closes with an outlook for future work and a summary in Section 6.

2. SOFTWARE ESCROW AS RISK MITIGATION

The risk of having dependencies to external third parties is not unusual to most industries outside of IT. However, most of the mitigation actions in real life simply change the relationship to the external partner by acquisition and integration into the own organisation. A recent study by Boston Consulting Group and UBS [3] indicates that nearly one in five of the companies surveyed intends to undertake at least one acquisition. At least 18% of the respondents stated "Access intellectual property and R&D" as main driver for M&A activities. So these acquisitions bypasses the risks introduced by outsourcing by changing the relationship to the third party.

The only technique that really mitigates the outsourcing risks while leaving the legal entity status of the outsourcing partner unaffected is outlined in the following subsections, followed by illustrating some pitfalls that motivate our improved approach.

2.1 Escrow Services

A well-established service to reduce the risks generated by strong dependencies to 3rd parties is to establish a so called "Escrow Service". A software escrow is a three-party arrangement, similar to a trust: "*An independent trustee – usually a firm in the business of doing technology escrows – is appointed is the escrow agent for licensor and licensee. The parties enter into a three-way agreement. The licensor delivers a copy of the source code to the escrow agent, and is usually required to deliver a source code update whenever it delivers a corresponding object code update to the licensee under the corresponding license agreement. Upon occurrence of a triggering event, and only then, the escrow agent delivers the escrowed source code to the licensee*". [12]

The risk mitigation approach is as follows: The software purchaser (i.e. the licensee) and the software provider (i.e. the licensor) maintain their legal status and even the level of information to be exchanged between both parties is unchanged. This is an important prerequisite to secure the intellectual property (IP) of the supplier.

In daily business the role of the trustee, the so-called Escrow Agent, does not affect the IP discussion as he receives all information (such as the source code) solely to file away. However, if a so called escrow clause is triggered (e.g. if the supplier goes bankrupt), and only then, the trustee hands out all information to enable the licensee (in the case of software escrow: the software purchaser) to enable the continued operation and maintenance of the licensed application.

This type of service is well established in todays IT market. Market leader NCC for example reports a revenue of 17,9m£ only in the UK with over 100 FTEs [14].

2.2 Pitfalls

During the worldwide financial crisis in 2009/10 some of our customers faced a scenario where the software escrow case occurred but the risks that should have been mitigated revealed their full impact as some key information stored in some digital artefacts were not available. The typical pitfalls around the established software escrow service can be classified into

- Missing artefacts: Software is more than only source code: *"A set of computer programs, procedures, and associated documentation concerned with the operation of a data processing system; e.g. compilers, library routines, manuals, and circuit diagrams."* [10] A software escrow service considering only the source code fails to account for the holistic nature of software. Nowadays a lot of implementation work is done outside the source code proper. Typical examples are models for code generation, architectural views, testware, technical documentation, used libraries, configurations of development environments etc. Without these additional digital artefacts the source code has only limited value: It cannot be understood, analysed, changed or outsourced to another vendor. The more complex and developed the applied technology (e.g., .Net or J2EE) the more business logic is stored in artefacts outside the source code.
- Low quality of deposited material: In many cases the deposited source code was either incomplete or inconsistent with the corresponding binary code. The source code was not commented, could not be analysed in any efficient way and did not follow standard software engineering techniques such as modularisation and decoupling. Exhuming a code basis with these attributes does not allow to re-compile/re-build the application and hinders any maintenance work that is necessary to adjust the application due to changed requirements.

If these pitfalls occur in real life the consequences can be devastating: It can start from the need for investing a large amount of money to conduct software-archaeology before continuing maintenance work and goes up to the complete re-development of the application being under software escrow.

2.3 Challenges

Consequently, the key challenge to be addressed in order to make the Escrow Service work in practice and avoid the aforementioned mistakes is to answer the following question:

How can we make sure that all relevant information for taking over an IT system exist and are of appropriate quality?

This is the point where we hope to bring in DP tools and concepts like [5]. Similar to software escrow services, DP tries to preserve digital artefacts necessary for assuring their availability over time. For software escrow, we need to preserve complete business processes (including tools, external knowledge etc.) at a sufficient level of quality of the preserved artefacts. This same is valid for DP (at least for digital objects), so the capability for reuse is obvious.

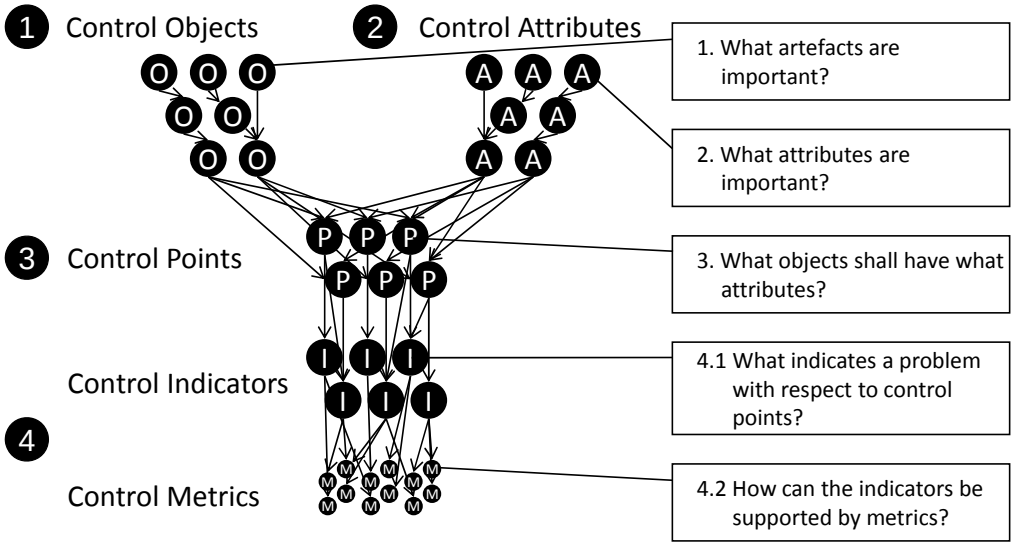


Figure 1: The QRM framework and its components

3. SOFTWARE ESCROW VIEW ON DP

3.1 How long is long-term?

Typically, DP is connoted with the aspect of long term preservation and most of the concepts of DP have been developed with the long term views (decades rather than months or years) in mind. We believe that the concepts developed so far are also very valuable in the case of software escrow and can be applied beneficially for much shorter periods of time. In some cases the timespan may only be a couple of months and we make use of a slightly different view on 'long-term'. Consider the definition for 'long-term' given by the OAIS:

"Long Term is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. Long Term may extend indefinitely." [5]

To our experience, Digital Preservation is not only required in the 'long-term' from a time based understanding as changes in technology can occur much more frequently. From the software escrow point of view involved parties have to keep information fit for purpose across the lifecycle of technologies (or any other kind of significant change in the context of the information that would normally render the respective information useless).

3.2 Basic Preservation Process in TIMBUS

According to TIMBUS project [19], one of the most up-to-date project funded by EU around Digital Preservation, the high level process of DP comprises three stages (cf. Figure 1).

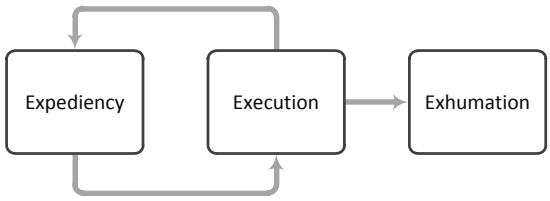


Figure 2: The three phases of the digital preservation process

- Expediency: In this, the fundamental steps need to be taken to determine what should be preserved.
- Execution: After the expediency has been established it is necessary to actually execute the DP preservation activities (e.g. conserving and archiving artefacts).
- Exhumation: In this stage, the preserved artefacts are brought back from the libraries into live environments to take up the regular 'business activities'.

Note that the stages and their activities are independent of the timespan of a preservation project, so it does not depend on the long-term view.

3.3 Models in DP

From our knowledge DP research so far has already taken care of "How to" preserve digital information by focussing on preservation processes and the lifecycle of different media, data formats and storage technologies (cf. [21], [22]). For DP in general (and software escrow in particular) the still open (but crucial) question is: What is the relevant information to be preserved and what digital objects (DO) contain this information? That is often not easy to answer because the boundaries of the system to be preserved are difficult to identify and usually debatable. For example, in almost all practical cases software depends on and makes use of third party components and libraries. To what extent do these third party artefacts need to be preserved? Where is the line between relevant context and the (for the time being) non-relevant context? This challenge increases when using Cloud services like SaaS or PaaS.

Our approach to answering the question of contexts to the best possible extent lies in using explicit models for the context of the systems to be preserved. The aim of these models is to preserve not only the DOs itself but additionally to capture the semantics of the objects. We make use of well-established architecture frameworks for specific domains as a starting point to identify and structure the DOs. Well known examples for these architecture frameworks are the NATO Architecture Framework (NAF) [19], the Zachmann framework [6] or The Open Group Architecture Framework (TOGAF) [20]. It is expected that these architecture models are describing a holistic view in their domain.

4. SOFTWARE ESCROW MODELLING APPROACH

The key to a successful Digital Preservation that can be completely used for software escrow is the holistic scrutiny of artefacts, their components and their respective properties. The vehicle we apply to fulfil this requirements of digtal objects is the so called quality risk management framework (QRM) [8]. The QRM framework has been used successfully as a foundation for project risk management and its concepts and ideas are applied in conjunction with DP concepts as to improve the success rate of software escrow.

4.1 Overview of the QRM Framework

The generic risk management framework consists of several components, whose instantiation is crucial for holistic software escrow. The overall QRM framework is depicted in Figure 2. The setup of the framework is explained in the following paragraphs.

Firstly, we identify the relevant DOs required for digital preservation by building a taxonomy for the context. In the framework these objects are named *control objects* (cf. Figure 2, "1"). Secondly, the quality attributes of digital objects are inventoried and classified – in terms of the QRM framework these are named *control attributes* (cf.Figure 2, "2"). In order to identify the essential aspects for digital preservation the Cartesian product of control objects and controls attributes is determined in a third step. The product of (control object, control attribute) is called a *control point* (cf. Figure 2, "3"). For each of the control points we determine its relevance on a Likert scale (e.g., ++ - very high, to -- – very low) indicating its priority for subsequent steps.

When control points are defined we have laid out the full view on what to preserve with which priority and which criticality. As a third step, we define *control indicators* and *control metrics* (cf. **Figure 2**, "4") supporting the control points with tangible information based on the artefacts.

How to apply these sequent steps in general? The solution is to reuse existing catalogues from other disciplines. For the development of the taxonomy (see above) the reuse of standard models for the relevant context is possible. For example, if there is a need to preserve the organisational context we can make use of a well-defined model such as the European Framework for Quality Management (EFQM) – model [15]. This model defines objects and their attributes for evaluating organisations and provides a valuable source for modelling the contexts to be preserved. A catalogue of preservable objects could be a taxonomy based on the EFQM-model which is illustrated in Figure 3.

But reuse can be done on the attribute level as well: they are independent from the objects in the first phase and can be derived from established standards such as ISO9126 [10], QUINT2 [17] and research in [2].

Both, the catalogue of digital objects and quality attributes are refined for the specific purpose of software escrow in the following Sections.

4.2 A Catalogue of Digital Objects for Escrow

In the software escrow use case for DP we need to ensure that the full set of digital objects required for the maintenance and evolution of a software system is preserved for all agreed releases of the software from the software provider's repositories. A first and simple approach to preservation of what is required in the software escrow case starts intuitively with the software's source

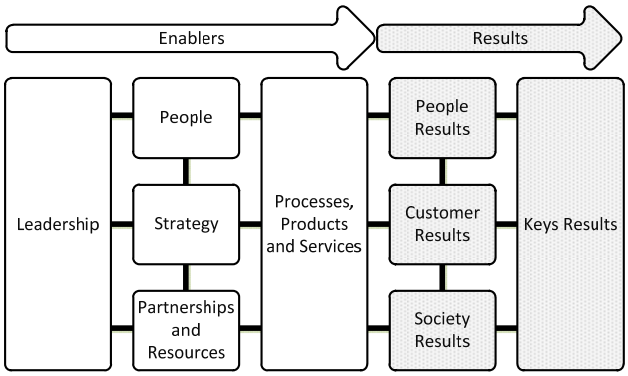


Figure 3: EFQM excellence model as initial context setting for software escrow in DP

codes. In case the software escrow partners are aware of the effort it takes to re-build the executable software system from the source code the compiled and ready-to-run executables are preserved additionally. To our experience, these types of digital objects are considered in the first place as it is one of the most obvious artefacts of value for a software purchaser.

However, this is by no means all it needs for a successful exhumation of the software at a later point in time. The IEEE Standard Glossary of Software Engineering Terminology reveals for good reasons a far wider definition of software [10], taking into account a far more holistic set of artefacts worth to be preserved.

A significant proportion of the artefact types and artefacts mentioned in [10] (compilers, library routines, manuals, documentation) is usually in practise not considered as a part of a software purchase and therefore tends to be neglected in the escrow preservation process. Examples for important documentation are the software architecture, the programmer's manual and other 'internal' documentation usually only required for maintenance purposes. (Which is exactly what the purchasing party wants to take over in case of the escrow exhumation.) More detailed taxonomies for documentation can be derived from text books such as Sommerville [17]:

- System Documentation
  - Requirements
  - System Architecture
  - Program Architecture
  - Component Description
  - Source-Code-How-To
  - Maintenance Guide
  - Environment Description
- End User Documentation
  - Functional Description
  - Reference Manual
  - Installation Manual
  - System administrators guide

In the case of software escrow exhumation, the software purchaser needs to take over the full maintenance process for the software under escrow. To be in a position to pick up these tasks in an efficient way, artefacts beyond the end user view are required. A first incomplete and project specific list contains

- configurations of the software and build environment
- the build environment itself and other third party tools and libraries
- software models and modelling tools

- test tools and test ware (tests, test data, automation, …)
- licenses to run the aforementioned tools and make use of 3rd party libraries
- licenses for intellectual property

Note that the escrow should ensure that for example licenses are issued for the purchaser, not for the original software developer. If license management is enforced by technical means it must be ensured licenses (and the depending tools) can be used for the purchaser.

Our current experience leads us to the taxonomy depicted in Figure 5. This taxonomy is usually used as a starting point for a more detailed elicitation and determination of the project specific DOs for software escrow. So far, we have seen a number of re-occurring DOs across different projects. But due to various reasons (e.g. software application domain terminology, business culture, or simply project lingo) it seems that the taxonomies are most useful if tailored to the project's context.

4.3 Quality Attributes of Digital Objects for Escrow

After having determined what to preserve for software escrow in the previous section, we address the properties of what to preserve in more detail. In many cases the exhumation already fails concerning a very simple attribute of the DOs – their existence. As many artefacts are forgotten or ignored the exhumation cannot be successful.

However, even if the artefacts do exist, the software purchasers must make their expectations towards the DOs explicit. If the purchaser has to take up maintenance activities they have to have an interest for example not only in the existence of relevant documentation but also in the quality of the respective documents. Thinking in terms of software engineering a good starting point for attributes of software artefacts is ISO 9126 [11] with QUINT2 [17] extensions (ISO 9126 has been superseded by the ISO 25000 series but remains a useful guidance for the purposes of this paper). Additionally, we enrich these attributes by attributes derived from research in the field of digital libraries [2]. The top level categories proposed can be reused by generalising their intended meaning from software to general artefacts. They are the following:

Control Attributes(ISO9126 and QUINT2, DL)			
Reliability		Maintainability	
	Maturity		Analyzability
	Fault tolerance		Changeability
	Recoverability		Stability
	Availability		Testability
	Degradability		Manageability
	Relevance		Reusability
	Significance		
Usability		Functionality	
	Learnability		Suitability
	Understandability		Accuracy
	Operability		Interoperability
	Explicitness		Compliance
	Customizability		Security
	Attractivity		Confidentiality
	Clarity		Integrity
	Helpfulness		Availability
	User-Friendliness		Traceability
	Accessibility		Completeness
	Consistency		Preservability
Portability		Efficiency	
	Installability		Time Behaviour
	Replaceability		Resource Behaviour
	Adaptability		Similarity
	Conformance		
	Timeliness		

Figure 4: Control attributes for digital objects in software escrow

- Reliability: A set of attributes that bear on the capability to maintain the level of performance under stated conditions for a presumed period.
- Usability: A set of attributes that bear on the effort needed for use, and on the individual assessment of such use, by a stated or implied set of users.
- Portability: A set of attributes that bear on the ability of artefacts to be transferred from one environment to another.

Digital Objects			
Interfaces	Documentation	Test Environment	Build Environment
Services Used	System	Infrastructure	Operating Systems
Data feeds	Requirements	Executables	Compilers
Network connections	System Architecture	Configurations	Programming Languages
Operating Systems	Program Architecture	Licenses	Runtime Environments
Services Provided	Component Description	Test cases	Configurations
Data feeds	Source-Code-How-To	Tests	3rd party libraries
Network connections	Maintenance Guide	Regression Tests	Licenses
Operating Systems	Environment Description	Test data	Applications
Business Processes	End User	Test scripts	Binaries
Intellectual Property	Functional Description	Automation	Executables
Patents	Reference Manual	Test reports	Database Schema
Algorithms	Installation Manual	Design Environment	Source Code
Methods	System administrators guide	Models	Configurations
Contracts	Processes	Configurations	
SLAs	Business Processes	Modelling tools	
UP	IT Processes	Licenses	
Supply Chains	Supporting Processes		

Figure 5: Digital objects for software escrow

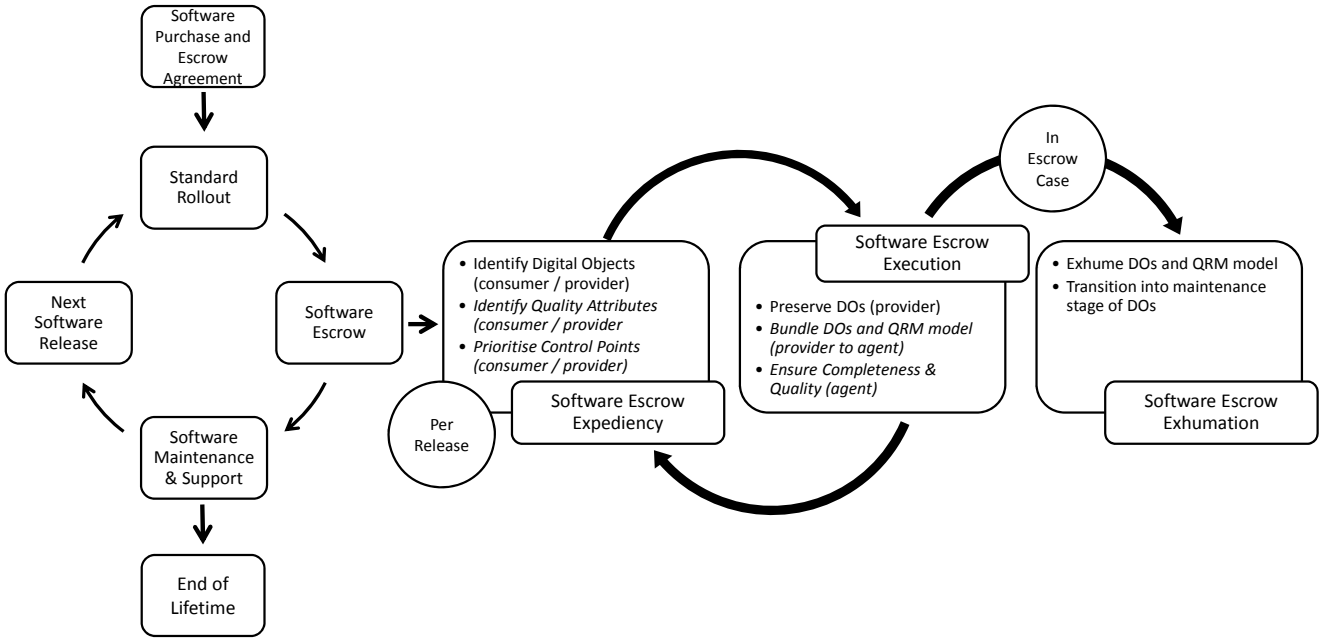


Figure 6: The high level software escrow process

- Maintainability: A set of attributes that bear on the effort needed to make specified (and consistent) modifications to artefacts.
- Functionality: A set of attributes that bear on the existence of a set of functions and their specified properties. The functions are those that satisfy stated or implied needs.
- Efficiency: A set of attributes that bear on the relationship between the level of performance of the software or the processes and the amount of resources used, under stated conditions.

The complete list of control attributes for software escrow is listed in Figure 4.

As before, the control attribute taxonomy needs to be tailored to the project context of the software escrow to be of most use. Having defined the taxonomy of control attributes independently from the specific control objects allows for an truly holistic view in the next step.

4.4 Software Escrow Control Points

Both the list of control objects and the list of desired control attributes can now be contrasted with each other. This can simply be done by calculating the Cartesian product of the two taxonomies and yields a matrix of all possible combinations of DOs with quality attributes. As we produce a full matrix we include potentially meaningless combinations of control objects and control attributes, we can now additionally prioritise the control points (e.g., on a scale from "not relevant" to "very important"). We can also use a more fine granular scale, but for illustrative purposes and in practical use in past projects already the simply five step scale proved effective. The prioritisation of the control points must be agreed between purchaser and vendor as it will guide the escrow agent in subsequent steps of the software escrow to assess both completeness and adequacy of the preserved DOs.

4.5 Indicators and Metrics

The sheer act of defining the control points itself already provides benefits as it clarifies what to look at and which attributes are relevant to which artefacts. In a final step, the control points are associated with *control indicators* and *control metrics* (cf. Figure 2, "4"). Control indicators help to identify quality risks by making use of simple metrics. The following example was arbitrary selected and only illustrates the idea in general. The control point ("Requirements", "complete"), being very important for both Digital Preservation and for software escrow, could be supported e.g. by an indicator "98% of requirements have an ID". The supporting metrics are (a) count requirements, (b) count requirements with ID, (c) compute the ratio of (a) and (b).

In practise, indicators are most successful when expressed in terms of non-desired properties. For example, it is very difficult to assess the quality of requirements written in natural language. Rather than trying to measure the quality directly, it is attempted to identify the "bad" requirements by searching for terms like "to do", "tbd", etc. If we identify one of the search terms in the context of a requirement (a task that can even be automated to some extent) we assume the requirement's quality is low: In this case it does not make sense to digitally preserve them nor does it make sense to be part of any software escrow.

By following this pattern of negating quality the set of indicators comprises a safety "net" of things we do not want to see. Having a sufficient number of indicators significantly reduces the risk of missing a bad "smell" and allows for re-adjustment of the quality model over time.

5. APPLYING DP IN SOFTWARE ESCROW USE CASES

The concepts described in the preceding sections are applicable to a multitude of use cases. In the following we will outline three of those use cases that highlight the value that the application of DP-techniques can add to software escrow.

5.1 Holistic Software Escrow

The overarching software escrow process starts when the two parties – software purchaser and software provider – agree the terms and conditions of the escrow contracts with the help of the software escrow agent. The software purchaser identifies the need for software escrow and subsequently both software purchaser and software provider prepare for an escrow agreement. The necessary steps for the execution in addition to the usual software rollout and maintenance procedures, and commitment on the triggers of the software escrow exhumation case are determined. The software escrow processes from the viewpoint of DP can be illustrated in . When comparing Figure 1 and Figure 6 it becomes obvious that these processes constitute a direct application of the DP processes to the software escrow problem space.

The first step, called software escrow expediency, aims at establishing (first iteration) or refining/revising (further software releases) the DOs, their quality attributes and the QRM model including the respective control points required. Secondly, per release of the software under escrow the software provider makes the preserved assets available to the escrow agent and bundles the DOs with their QRM model. The escrow agent then can ensure the completeness and adequate quality of the artefacts provided by the software provider (utilising technical support) without disclosing the preserved information to the software purchaser. If the escrow case does not occur before the next release of the software product the escrow process is hibernated. The agreed software maintenance and support is delivered by the provider. Typically, software products are updated from time to time and consequently, after every new rollout of a new release of the software under escrow the escrow process is triggered again.

In case the predefined events terminating the existence of the software provider occur, escrow exhumation is triggered. The software escrow agent hands over to the software purchaser all assets in his behold. The software purchaser then may take the necessary step as to re-vitalise the software maintenance and support activities either on his own or with the support of a different software supplier.

This approach is well-established in the software and IT industry and there is a variety of vendors that offer this software escrow service. Figure 7 illustrates the roles and tasks for the software escrow approach, in which the software provider hands over certain assets being part of their intellectual property to an escrow agent who safely files away and manages access to those assets. The software purchaser simply gets the executable software, just as would be the case without the software escrow agreement.

Usually, the assets would remain in possession of the escrow agent until the contract between software provider and software purchaser terminates or until other events render the escrow unnecessary.

Unfortunately, with software escrow, risks come into effect at a late point in time: If and only if the escrow event occurs the software purchaser will get access to the escrow asset base, and only then is he able to determine the suitability of stored assets.

The holistic software escrow, utilising the Digital Preservation research area, goes beyond the provisioning of a simple storage and management service by the escrow agent by utilizing the power of the approach detailed in the preceding sections (cf. Section 4). By including an appropriate amount of quality assurance into the escrow process upon software escrow expediency and software escrow execution, the rate of success upon software escrow exhumation can be greatly increased, increasing trust with

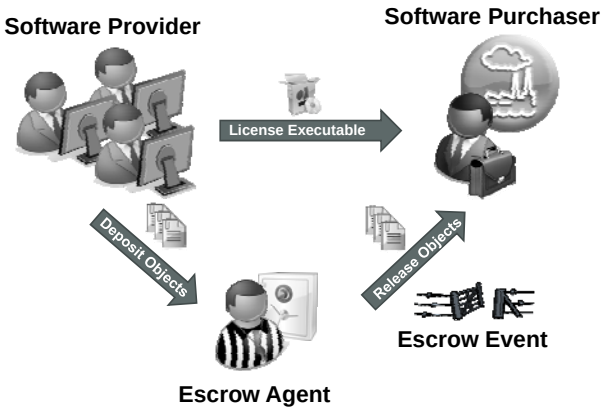


Figure 7: Roles and tasks of the software escrow scenario

the software purchaser and enabling the software provider to ask for higher compensation for the software escrow option.

5.2 Ex-Post ESCROW Analysis

As discussed previously, there already exists a market for software escrow that has developed primarily in the Anglo-Saxon countries, where players are offering a fairly basic escrow service with limited success, as more often than not the quality of deposited assets is insufficient for exhumation of the software at a later point in time.

Subsequently, there is a need for the software purchaser to determine the course of action in such a situation. The fundamental question that needs answering is whether it is worthwhile to invest into re-engineering the system based on its available artefacts or whether the system needs to be rebuilt from scratch, discarding whatever was supplied as part of the escrow effort.

The procedure of choice for evaluating the various alternatives and for answering this underlying question is to estimate the respective investments for the relevant alternatives. For the alternatives that target the re-use of assets from the escrow this requires transparency about the status quo as well as enough data to support a reliable estimation of effort necessary to transform those assets into value for the business.

An ex-post escrow analysis as depicted in Figure 8 yields the necessary transparency by

- identifying the software purchaser's vision, goals and subsequent requirements towards the system
- making use of the software escrow object catalogue to define a target state of required assets
- making use of the software escrow attribute catalogue to map the software purchaser's requirements to and prioritize attributes
- conducting a gap analysis to identify the gap between this target state and the status quo
- estimating the effort required to fill this gap by reverse engineering of those parts of the system that are missing, incomplete, or outdated

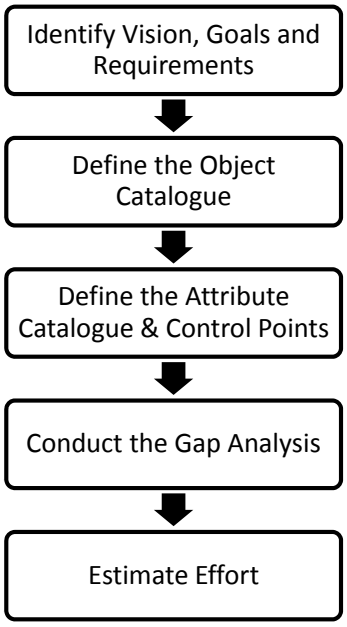


Figure 8: The ex-post software escrow process

When comparing this list of activities with the tool box supplied by DP, it quickly becomes evident that we can draw heavily on DP techniques to conduct the abovementioned tasks. The following sections detail this link.

5.2.1.1 Identifying Vision, Goals and Requirements

Before any of the alternatives can be evaluated and ranked against each other, a precise definition of the target state is required. The required documentation to define this target state consists of strategic visions, broken down into goals and objectives which in turn decompose into a set of requirements that operationalise these goals (similar to the first steps in the GQM approach [2]).

Some of this information will be available already, while others may not. In any case, existing documentation needs to be reviewed to gauge its accuracy and actuality before it may be used as a foundation for the analysis. Any documentation that does not yet exist needs to be made explicit. The discipline of Requirements Engineering has developed a proven set of methods to extract these requirements using a variety of techniques [15].

5.2.1.2 Defining the Object Catalogue

Once the high level target state is known, the preparation of the actual gap analysis can commence. The first step towards this gap analysis comprises the tailoring of the escrow object catalogue to fit the specific requirements in place (cf.Figure 2, "1"). The standard escrow object catalogue depicted in Figure 5 is used as a starting point from which all irrelevant objects are stripped.

The result of this activity comprises a catalogue of objects that need to be present within the software escrow asset base for the assets to be used in the intended fashion. This catalogue constitutes the starting point for setting up an escrow quality model that will be used to support the subsequent gap analysis.

5.2.1.3 Defining the Attribute Catalogue & Control Points

More often than not the sheer existence of an artefact is not enough to render it a useful asset that supports the viability of a system.

Subsequently, each of the objects in the object catalogue must be part of an overall QRM model and needs to be defined that establishes and operationalizes the expected quality for this specific object (cf. Figure 2, "2"). The generic escrow attribute catalogue depicted in Figure 4 serves as starting point for the analysis of each escrow object that determines how the previously identified requirements apply to each of the escrow objects (cf. Figure 2, "3").

These combinations of escrow objects with escrow attributes are result in the escrow control points. For the above mentioned reasons, not all combinatorial possible combinations of escrow objects and escrow attributes are meaningful. The meaningful ones to be considered require a prioritization due to economic reason, as discussed in Section 4.4.

Additionally, for each such escrow control point, a suitable verification method needs to be determined and documented, including all parameters that may have an influence on the result of the verification. The verification method can be an in-depth analysis of the control object with regards to the respective control attribute in the most complex case or, in simpler cases can be supported by – simple – indicators and metrics.

5.2.1.4 Conducting the Gap Analysis

The actual execution of those verification activities involves the application of methods and techniques from the quality assurance discipline to the software escrow asset base in order to determine the degree of gap that may exist between target state and the artefacts contained in the asset base.

The results of these verification activities are mapped to the relevant software escrow control points (cf. Figure 2, "4"). Doing this guarantees traceability from verification results back to individual escrow objects and attributes as well as the ability to aggregate the results.

Once verification activities have concluded, the so annotated QRM model is used to systematically identify the gaps between verification results and target state.

5.2.1.5 Estimating Effort

Finally, the gap analysis results are used to inform the effort estimation that makes the cost of using the escrow asset base explicit by attaching a figure to it.

For each of the gaps identified, viable mitigations need to be identified and for each of those expected effort and cost needs to be estimated. In addition to this, all other direct and indirect costs, such as costs arising from the need to license third party intellectual property in order to use the escrow asset base need to be considered. This holistic estimate of costs associated with (re-)use of the escrow asset based can then be used as part of a larger evaluation and decision making process that ranks all potential alternatives against each other using the predefined requirements.

In supporting this decision making, the ex-post escrow analysis can contribute as much value to the business as is possible for a post-mortem analysis. While it is certainly able to create transparency regarding the viability of the deposited escrow asset base it cannot bring back artefacts that have not been deposited, be it intentionally or for want of knowledge that certain artefacts are required. In a worst case scenario, the ex-post analysis can only establish that the deposited assets are without any value to the software purchaser and thus do not constitute a viable alternative. The ability to influence the course of action before any damage is done is a luxury that is only afforded to the holistic software

escrow documented in Section 5.1 and utilising the Digital Preservation knowledge base.

5.3 Due diligence Analysis

Mergers and acquisitions (M&A) of companies are a risky undertaking. Recent studies find that almost two thirds of all mergers and acquisitions fail, for instance resulting in a split along old corporate borders [9].

In those cases where the corporate management needs to report to a diverse group of owners and other stakeholders such as with publicly traded companies there is a subsequent requirement to prove to owners and stakeholders that corporate management is diligent in executing the merger or acquisition by closely scrutinizing the partner or acquiree. One integral part of this scrutiny is a financial valuation of the organization and all its assets.

IT systems and the software and applications that drive those systems are usually part of the tangible assets that are owned by any modern company. Unlike real estate, a corporate fleet or factory buildings with production lines inside, software is notoriously difficult to value correctly. In addition to this, more complex M&A scenarios may require parts of the affected companies and their assets to be severed from the rest of the organization, for instance to be sold off separately because of regulatory concerns.

All this calls for both the precise valuation of software and IT assets in general and for the ability to safely deposit assets into escrow while the M&A transactions are being finalized by all affected parties.

Subsequently, this use case constitutes a hybrid between a software escrow, where software assets are being put into Escrow and an ex-post Escrow analysis where an extant asset base is evaluated in terms of its future viability for the intended use cases.

This dichotomy becomes transparent when inspecting the methodological building blocks necessary to conduct this analysis:

- In a first step all IT and software assets relevant for the analysis need to be surveyed and mapped to an overall IT and software landscape that shall serve as input to the valuation. In terms of the QRM framework, the establishment of the control objects catalogue is a useful tool to achieve a comprehensive overview of existing assets.
- The valuation of assets itself can be conducted using the procedure detailed in Section 5.2, with the only real difference being that the assets to be analysed are not part of an software escrow asset base. This requires additional preparatory activities to collect and collate all the required assets for the analysis. From the QRM framework perspective, this step corresponds to the establishment of the control attributes catalogue and the subsequent elicitation of control point.
- Once the asset base is complete and the ex-post analysis has been conducted, a first value estimate can be delivered. The estimation can make use of the QRM model by supporting the various control points with price indicators and thereby systematically derive a transparent overall assessment.
- Depending on the discussed influencing factors, some assets may need to be put into software escrow for the duration of time during which the merger or acquisition is being executed.
- In order to diligently curb risk, all parties involved need to ensure that a holistic software escrow is instituted to make sure that whatever is put into escrow conforms to its estimated value after the merger or acquisition has been finalized.

6. SUMMARY AND OUTLOOK

In this paper we have laid out use cases taken from the industry scenario "IT outsourcing" that mitigate specific risks by making use of software escrow services and due diligence analyses. To support the use cases in practise we exploit the concepts of DP and apply the QRM framework to provide an holistic view on quality risk management. We have pointed out that the results from DP research are not limited to long-term views but can also be deployed in typically short-term scenarios.

Derived from our experiences we suggest a new definition of a holistic software escrow based on the definition made in [12]: "*An independent trustee is appointed as the escrow agent for licensor and licensee. The parties enter into a three-way agreement. The licensor delivers a copy of* ***all source artefacts needed to build the object code and maintain the software*** *to the escrow agent, and is usually required to deliver a update of the* ***artefacts*** *whenever it delivers a corresponding object code update to the licensee under the corresponding license agreement. Upon occurrence of a triggering event, and only then, the escrow agent delivers the escrowed artefacts to the licensee"*.

Currently, we are evaluating the suggested definition and further use cases –in the context of the TIMUBS project [19] – for applying QRM and DP in various contexts for the benefit of our customers.

7. ACKNOWLEDGEMENTS

Parts of this work have been supported by the European Union in the TIMBUS project [19]: "Digital Preservation for Timeless Business Processes and Services", Grant Agreement Number 269940

8. REFERENCES

[1] American Society for Quality/ISO 8402:1994
[2] Basili, V., Caldiera, G., Rombach, H. D.: The Goal Question Metric Approach. In: Encyclopedia of Software Engineering. John Wiley & Sons, 1994
[3] BCG The Boston Consulting Group: "M&A: Ready for Liftoff? A survey of European Companies' Merger and Acquisition Plans for 2010", December 2009
[4] Beagrie, N., Jones M. (maintained by Digital Preservation Coalition): Preservation Management of Digital Materials:

The Handbook, 2008, available at http://www.dpconline.org/advice/preservationhandbook
[5] Consultative Committee for Space Data Systems: Reference model for an Open Archival Information System recommendation for space data system standards. Washington D.C, 2002
[6] Zachmann, J.A.: A framework for information systems architecture, IBM Systems Journal 26 (1987)
[7] Goncalves, M.A., Moreira, B.L., Fox, E.A., Watson, T.L.: ''What is a good digital library?'' – A quality model for digital libraries, Information Processing and Management 43 (2007)
[8] EFQM Excellence Model 2010. http://www.efqm.org/
[9] Fulmer, R.M., Gilkey, R.: Blending Corporate Families: Management and Organization Development in a Postmerger Environment, The Academy of Management Executive Vol. 2, No. 4 (Nov., 1989)
[10] IEEE Standard Glossary of Software Engineering Technology, 1983
[11] International Standard ISO/IEC 9126, Part 1, Software engineering – Product quality – Quality model, Beuth-Verlag, Berlin (2001)
[12] Meeker, H.: "Thinking outside of the Lock Box: Negotiating Technology Escrow", Computer & Internet Lawyer, No 9, 2003
[13] NATO Consultation, Command and Control Board, NATO Architecture Framework (NAF), http://www.nhqc3s.nato.int
[14] NCC Group plc: "Preliminary Annual Results for the year ended 31 May 2010", July 2010
[15] Pohl, K.: Requirements Engineering, Springer, 2010
[16] Simon, F., Simon, D.: Qualitäts-Risiko-Management, Logos Verlag, November 2010
[17] Software Engineering Research Centre Netherlands, "Kwaliteit van Softwareprodukten – Ervaringen met een kwaliteitsmodel", http://www.serc.nl/quint-book.
[18] Sommerville, I.: Software Engineering, 8th edition, Pearson Education (2007)
[19] TIMBUS, http://www.timbusproject.net/
[20] The Open Group Architecture Framework, http://www.opengroup.org/togaf
[21] PLATO, The Preservation Planning Tool, http://www.ifs.tuwien.ac.at/dp/plato
[22] H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, M. Jehn, "nestor-Handbuch: Eine kleine Enzyklopädie der Langzeitarchivierung", 2009

New Dimension in Relational Database Preservation: rising the abstraction level

Ricardo André Pereira Freitas
CLEGI – Lusiada University
Vila Nova de Famalicão – Portugal
freitas@fam.ulusiada.pt

José Carlos Ramalho
Department of Informatics – University of Minho
Braga – Portugal
jcr@di.uminho.pt

ABSTRACT

The work addressed in this paper focuses on the preservation of the conceptual model within a specific class of digital objects: Relational Databases. Previously, a neutral format was adopted to pursue the goal of platform independence and to achieve a standard format in the digital preservation of relational databases, both data and structure (logical model). Currently, in this project, we address the preservation of relational databases by focusing on the conceptual model of the database, considering the database semantics as an important preservation "property". For the representation of this higher layer of abstraction present in databases we use an ontology based approach. At this higher abstraction level exists inherent Knowledge associated to the database semantics that we tentatively represent using "Web Ontology Language" (OWL). We developed a prototype (supported by case study) and define a mapping algorithm for the conversion between the database and OWL. The ontology approach is adopted to formalize the knowledge associated to the conceptual model of the database and also a methodology to create an abstract representation of it.

Keywords

Digital Preservation, Relational Databases, Ontology, Conceptual Models, Knowledge, XML, Digital Objects

1. INTRODUCTION

In the current paradigm of information society more than one hundred exabytes of data are used to support information systems worldwide [1]. The evolution of the hardware and software industry causes that progressively more of the intellectual and business information are stored in computer platforms. The main issue lies exactly within these platforms. If in the past there was no need of mediators to understand the analogical artifacts today, in order to understand digital objects, we depend on those mediators (computer platforms).

iPRES2011, Nov. 1–4, 2011, Singapore.

Our work addresses this issue of Digital Preservation and focuses on a specific class of digital objects: Relational Databases (RDBs). These kinds of archives are important to several organizations (they can justify their activities and characterize the organization itself) and are virtually in the base of all dynamic content in the Web.

In previous work [2] we adopted an approach that combines two strategies and uses a third technique — migration and normalization with refreshment:

- Migration which is carried in order to transform the original database into the new format – Database Markup Language (DBML) [3];
- Normalization reduces the preservation spectrum to only one format;
- Refreshment consists on ensuring that the archive is using media appropriate to the hardware in usage throughout preservation [4].

This previous approach deals with the preservation of the Data and Structure of the database, i.e., the preservation of the database logical model. We developed a prototype that separates the data from its specific database management environment (DBMS). The prototype follows the Open Archival Information System (OAIS) [5] reference model and uses DBML neutral format for the representation of both data and structure (schema) of the database.

1.1 Conceptual Preservation

In this paper, we address the preservation of relational databases by focusing on the conceptual model of the database (the information system – IS). It is intended to raise the representation level of the database up to the conceptual model and preserve this representation. For the representation of this higher level of abstraction on databases we use an ontology based approach. At this level there is an inherent Knowledge associated to the database semantics that we represent using OWL [6]. We developed a prototype (supported by case study) and established an algorithm that enables the mapping process between the database and OWL.

In the following section, we overview the problem of digital preservation, referring to the digital object, preservation strategies and the preservation of relational databases. Section 3 describes our previous work and states the open issue

(database semantic representation) the lead us to the current approach. In Section 4 we outline the relation between ontologies and databases establishing the state-of-the-art and referring to related work. The prototype and the mapping process from RDBs to OWL is detailed in section 5. At the end we draw some conclusions and specify some of the future work.

2. DIGITAL PRESERVATION

A set of processes or activities that take place in order to preserve a certain object (digital) addressing its relevant properties, is one of the several definitions. Digital objects have several associated aspects (characteristics or properties) that we should consider whether or not to preserve. The designated community plays an important role and helps to define

> *"The characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record"*[7].

2.1 The Digital Object

Some distinction can be established between digital objects that already born in a digital context, and those that appear from the process of digitization: analog to digital. In a comprehensive way and encompassing both cases above, we can consider that a digital object is characterized by being represented by multiple bitstreams, i.e., by sequences of binary digits (zeros and ones).

We can question if the physical structure of the object (original system) is important, and if so, think about possible strategies for preservation at that level, e.g. "technology preservation" (museums of technology) [8]. Nevertheless, the next layer — the logical structure or logical object—, which corresponds to the string of binary digits have different preservation strategies. The bitstream have a certain distribution that will define the format of the object, depending on the software that will interpret it. The interpretation by the software, of the logical object, provides the appearance of the conceptual object, that the human being is able to understand (interpret) and experiment. The strategy of preservation is related to the level of abstraction considered important for the preservation [9]. From a human perspective one can say that what is important to preserve is the conceptual object (the one that the humans are able to interpret). Other strategies defend that what should be preserved is the original bitstream (logical object) or even the original media. Figure 1 shows the relationship between the different levels of abstraction (digital object) and the correspond preservation formats adopted for RDBs in this research.

2.2 Relational Databases Preservation

By focusing on a specific class or family of digital objects (relational databases), questions emerge such as: what are the effects of cutting/extracting the object from its original context? Can we do this even when we are referring to objects that are platform (hardware/software) dependent? The interaction between the source of the digital object and

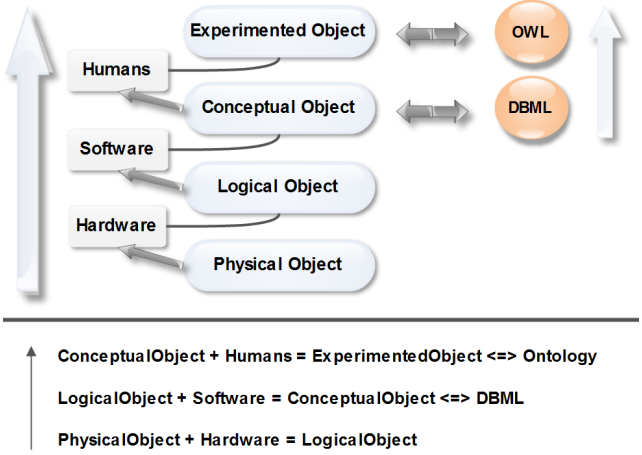


Figure 1: Levels of Abstraction and Preservation Policy

the platform results on a conceptual object that can be different if the platform changes [7]; the output can be different (will the object maintain its original behavior?). The important is the preservation of the essential parts that purport what the object where made for. Either the source or the platform can be altered if what is essential is obtained and also maintaining the meaning of the digital object over long periods of time (long-term scope).

Considering the nature of the digital artifacts that we are addressing – relational databases – there is an European strategy encompassed in the "Planets Project" [10] to enable their long term access. The project adopted the SIARD [11] solution, which is based on the migration of database into a normalized format (XML – eXtensible Markup Language [12]). The SIARD was initially developed by the Swiss Federal Archives (SFA).

Another approach, also based on XML, relies on the main concept of "extensibility" – XML allows the creation of other languages [13] (it can be called as a meta language). The DBML [3] (Database Markup Language) was created in order to enable representation of both **DATA** and **STRUCTURE** of the database.

Both approaches (SIARD and DBML) adopt the strategy of Migration of the database to XML, why? A neutral format that is hardware and software (platform) independent is the key to achieve a standard format to use in digital preservation of relational databases. This neutral format should meet all the requirements established by the designated community of interest.

3. PREVIOUS WORK AND CURRENT APPROACH

In previous work we address the preservation of the RDBs data and structure by developing a archive prototype that uses the DBML format for preservation. Our first approach covers the preservation of the logical model of databases (tables, structure and data). However, neither this approach nor others (e.g. SIARD [11]) is concerned with the database

Digital Object	Preservation Levels	Relational Database
Experimented Object	**Ontology**	**Conceptual Model**
Conceptual Object	DBML	Logical Model
Logical Object	–	Original Bitstream
Physical Object	–	Physical Media

Figure 2: Preservation Policy

semantics. The focus of our research then turned into this problem related conceptual model of the database, i.e., the information system on the top of the operational database.

3.1 First Approach

The prototype is based on a web application with multiple interfaces. These interfaces have the mission to take a certain database and ingest it into the archive. The access to the archive in order to do all the necessary interventions on the system is also done through those web interfaces.

Conceptually, the prototype is based on the OAIS [5] reference model. The OAIS model of reference does not impose rigidity with regard to implementation, rather it defines a series of recommendations. The OAIS model is accepted and referenced for digital preservation purposes since it is concerned about a number of issues related to preservation of digital artifacts: the process of information Ingestion into the system, the information storage as well as its administration and preservation, and finally information access and dissemination [14] [15]. Three information packages are the base of the archival process: Submission Information Package (SIP), Archival Information Package (AIP) and Dissemination Information Package (DIP).

3.2 Proposed Approach

Based on the first prototype we now intend to include in the information packages (SIP, AIP and DIP) an higher representation level of the database — the conceptual model of the database. Ontologies are used to address semantics and conceptual model representation.

Our hypothesis concentrates on the potentiality of reaching relevant stages of preservation by using ontologies to preserve of RDBs. This lead us to the preservation of the higher abstraction level present in the digital object, which corresponds to the database conceptual model. At this level there is an inherent **Knowledge** associated to the database semantics (Fig. 2).

We intend to capture the experimented object (knowledge) through an ontology based approach. This experimented or knowledge object is the "final abstraction". The ontology approach is adopted to formalize the knowledge present at the experimented object level and also a methodology to create an abstract representation of it. The system has evolved into an OAIS based architecture that allows the ingestion, preservation and dissemination of relational databases at two levels of abstraction — logical and conceptual (Fig. 2). This approach is also an extension to previous approaches in terms of metadata since the ontology provides information about the data at a conceptual level. Figure 2 also shows a possible preservation "lifecycle" of RBDs.

4. ONTOLOGIES & DATABASES

There is a direct relation between ontologies and databases: a database has a defined scope and intends to model reality within that domain for computing (even when it is only virtual or on the web); ontology in ancient and philosophical significance means the study of being, of what exists [16].

The (strong) entities present in relational databases have an existence because they were model from the real world: they relate to each other and have associated attributes. In information society and computer science, an ontology establishes concepts, their properties and the relationships among them within a given domain [17].

4.1 Database Semantics

A database can be defined as a structured set of information. In computing, a database is supported by a particular program or software, usually called the Database Management System (DBMS), which handles the storage and management of the data. In its essence a database involves the existence of a set of records of data. Normally these records give support to the organization information system; either at an operational (transactions) level or at other levels (decision support – data warehousing systems). In particular, the relational databases model is designed to support an information system at its operational level. Thus, RDBs are complex and their data can be distributed into several entity relations that related to each other through specific attributes (foreign to primary keys) in order to avoid redundancy and maintain consistency [18].

If we intend not only to preserve the data but also the structure of the (organization) information system we should endorse efforts to characterize (read) the database semantics. It is intended to raise the representation level of the database up to the conceptual model and preserve this representation. In other words, we represent the conceptual model of the database using an ontology for preservation.

4.2 Ontologies

The study of ontologies in computer science received new impetus due to the growth of the web, their associated semantics and the possibility of extracting knowledge from it. Tim Berners-Lee realized that years ago giving origin to the "Semantic Web" supported by W3C (World Wide Web Consortium) [19] which works on establishing a technology to support the *Web of data* [20]. Notice that a tremendous part of the web is based in (relational) databases — specially dynamic information.

Behind the ontology there is the need of knowledge representation for machine interpretation. Two technologies: a) the RDF (Resource Description Framework) [21] triples give support for the meaning in simple sentences b) and XML [12] is used for structuring documents [16]. The RDF document consist on a set of triples, – *object, property, value* – that we can also define as – *subject, predicate, object* [22].

The notion of ontology then emerges due to the need of expressing concepts in different domains (ontologies as collections of information). An ontology can provide readable information to machines [23] at a conceptual level (higher

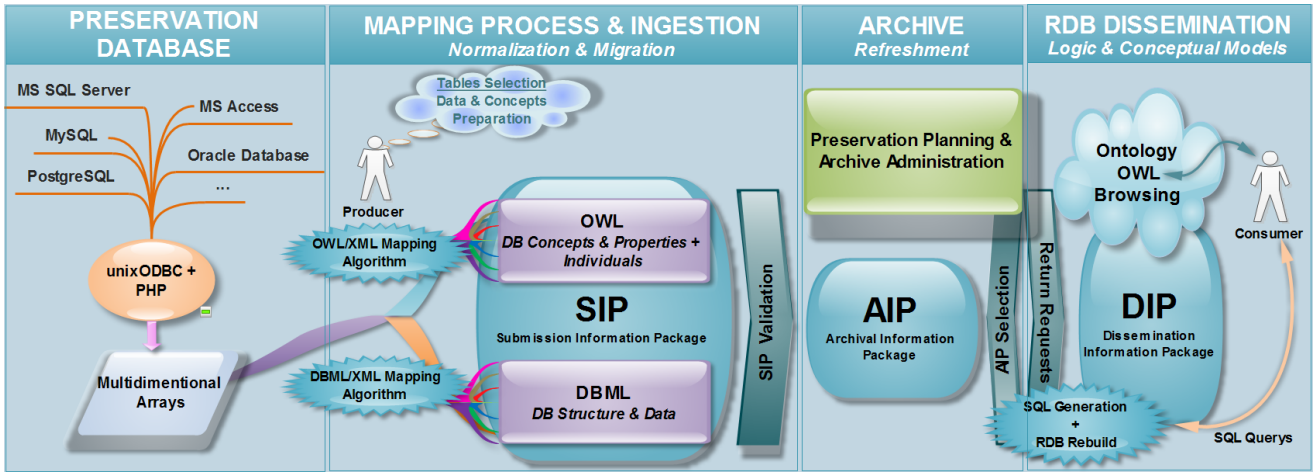


Figure 3: RDBs Preservation Framework

abstraction level). They also enable the integration and interpretability of data/information between applications and platforms. Ontologies benefit from the fact that they are not platform/system dependent when compared to traditional relational databases.

4.3 Related Work

Work related to RBDs and ontologies transformations proliferate and is addressed continuously. Considering the RDF [21], OWL [6], ontologies and RDBs, several frameworks, mapping approaches and tools exist: Virtuoso RDF View [24]; D2RQ [25]; Triplify [26]; RDBToOnto [27]; R2O [28]; Dartagrid Semantic Web toolkit [29]; SBRD Automapper [30]; XTR-RTO [31]; RDB2OWL [32]; DB2OWL [33]; R2RML [34]; OntER [35]; DM2OWL [36]; OWLFromDB [37] and also "Concept hierarchy as background knowledge" proposal [23] among others.

Several of these approaches and tools are referenced and analyzed in the W3C Incubator group survey [38] and also in [23].

The conversion from databases into an ontology could be characterized as a process in the scope of reverse engineering [35]. While some approaches and works try to establish a mapping language or a mapping process [39], others use different techniques and strategies for the database translation [36] into an ontology (e.g. OWL).

The R2RML (RDB to RDF Mapping Language) [34] working draft submitted to W3C is designed for mapping the data within the attributes of a **table** into pairs: property, object. Each record within a table share the same subject in this RDF triple map relation. This approach supports the input of "logical" tables from the source database, which can be an existing table, a view or a valid SQL query. Also in cases where attributes are foreign keys it is generated a pair (property, object) referencing the correspondent table. The rules for this mapping are then organized in a vocabulary with several classes and subclasses (*TripleMapClass, SubjectMapClass, PredicateMapClass, ObjectMapClass, RefPredicateMapClass, etc*).

For example, R2O [28] approach is based on a mapping document generation (mapping language). Virtuoso RDF View establishes a set of RDF statements by defining for each table: *primary key* (subject), *attribute* (predicate), *value* (object). In the RDB2OWL [32] a different strategy is used since it is created a mapping RDB schema. The "Concept hierarchy as background knowledge" proposal [23] gives special attention to the data preparation before conversion and to the knowledge that resides on the database.

5. FROM RDB TO OWL

This section presents the work developed to convert databases to ontology, based on a mapping process (mapping algorithm), for preservation. We intend to preserve a snapshot of the database (or a frozen database) by preserving the OWL generated from the database.

We start by concentrating our efforts on detailing the mapping process and analyzing the created algorithm. Then the conducted tests and some of the results are also presented.

5.1 Mapping Process of RDBs to OWL – Prototype

Our work implements the conversion from RDBs into OWL through an algorithm that performs the mapping process. The developed prototype enables the connection to a DSN (Data Source Name), extracts the data/information needed and gives the initial possibility of selecting the tables of interest (for conversion). It is assumed that the source database is normalized (3NF).

Lets start by enumerating the properties of RDBs that are address and incorporated in the ontology (OWL):

- **Tables** names;
- **Attributes** names and data types;
- **Keys** primary keys, foreign keys (relationships between tables);
- **Tuples** data;

```
tables = Array{ [1] => t1, ... , [n]=> tn }

columns = Array{
                [t1] => Array{
                        [a1] => Array{ [Name] => 'a1_name', [Type] => 'a1_type' },
                        ...,
                        [an] => Array{ ... }},
                        ...,
                [tn] => Array{...}}

p_keys = Array{
                [t1] => Array{ [a1] => 'pk_t1', ..., [an] => 'pk_t1' },
                ...,
                [tn] => Array{...}}

f_key = Array{
                [t1] => Array{
                        [a1] => Array{ [pk_table] => 'tref', [pk_column] => 'tref.aref'},
                        ...,
                        [an] => Array{...}},
                        ...,
                [tn] => Array{...}}

tables_data= Array{
                [t1] => Array{
                        [1] => Array{ [a1]=> 'a1_data', ..., [an] => 'an_data'},
                        ...,
                        [m] => Array{...}},
                        ...,
                [tn] => Array{...}}
```

Figure 4: Multidimensional Array Structure

```
//   Classes (tables) & ObjectProperties (link tables - non_classes)
FOREACH [ table ]
        IF [ ( |columns[table]| = |p_keys[table]| ) AND ( |p_keys[table]| = |f_keys[table])| ) ] THEN
                non_class[] = table
                FOREACH [ columns[table] - 1 ]

                NEW 'ObjectProperty'
                        Property_Description = 'is_' + f_keys[table][columns[table]][pk_table] + '_Of'
                        Domain = f_keys[table][columns[table]][pk_table]
                        Range = f_keys[table][next(columns[table])][pk_table]

                NEW 'ObjectProperty'
                        Property_Description = 'has_' + f_keys[table][columns[table]][pk_table]
                        Domain = f_keys[table][next(columns[table])][pk_table]
                        Range = f_keys[table][columns[table]][pk_table]

                NEW 'InverseObjectProperties'
                        Property_Description = 'is_' + f_keys[table][columns[table]][pk_table] + '_Of'
                        Property_Description = 'has_' + f_keys[table][columns[table]][pk_table]

                END FOR
        ELSE
        class[] = table
        END IF
END FOR
```

Figure 5: Algorithm – Classes and Non Classes

These elements are extracted from the database into multidimensional arrays. Figure 4 shows the arrays structure.

For each `table` on the database we define a **class** on the ontology with the exception of those tables where all attributes constitute a composed primary key (combination of foreign keys). These link tables used in the relational model to dismount a many-to-many relationship, are not mapped to OWL classes, instead they give origin to **object properties** in the ontology. These object properties have on their domain and range the correspondent classes (database tables) involved in the relationship (Fig. 5).

The `foreign keys` of the tables mapped directly to OWL classes also give origin to **object properties** of the correspondent OWL classes (tables). The `attributes` of the several tables are mapped to **data properties** within the analogous OWL classes with the exception of the attributes that are foreign keys (Fig. 6).

The algorithm generates inverse object properties for all relationships among the classes. If the object properties are generated directly from a 1-to-many relationship (which is the last case) it is possible to define one of the object properties as functional (in one direction).

The `tuples` of the different tables are mapped to **individuals** in the ontology and are identified by the associated `primary key` in the database. A tuple in a database table is mapped to an individual of a class (Fig. 7).

The object properties that relates individuals in different

```
//   Sub Classes of Thing & Disjoint all & Object and Data Properties
class_disjoint[] = class
FOREACH [ class ]
        NEW class 'SubClassOf' owl:Thing
        FOREACH [ class_disjoint ]
                IF  [ class IN class_disjoint ] THEN
                        NEW 'DisjointClasses'
                                Class_Description = class
                                Class_Description = class_disjoint
                END IF
        END FOR
        pop(class_disjoint)

        FOREACH [ f_keys[table] as fk ]
                NEW 'ObjectProperty'
                        Property_Description = 'is_' + fk['pk_table'] + '_Of'
                        Domain = fk['pk_table']
                        Range = class
                NEW 'objectProperty'
                        Property_Description = 'has_' + fk['pk_table']
                        Domain = class
                        Range = fk['pk_table']
                NEW 'InverseobjectProperties'
                        Property_Description = 'is_' + fk['pk_table'] + '_Of'
                        Property_Description = 'has_' + fk['pk_table']
                NEW 'FunctionalobjectProperty'
                        Property_Description = 'is_' + fk['pk_table'] + '_Of'
        END FOR

        FOREACH [ columns[table] as table_data ]
                IF [ f_keys[$table][table_data['Name']]['pk_column'] != table_data['Name'] ] THEN
                        NEW 'DataProperty'
                                Property_Description = 'has_' + table_data['Name']
                                Domain = class
                                Range = data_type
                END if
        END FOR
END FOR
```

Figure 6: Algorithm – Structure Generation

```
//   tuples -> Individuals   //
FOREACH [ class ]
        FOREACH [ tables_data[table] as tuple ]
                primary_key = class
                FOREACH [ p_keys[table] as pk)
                        primary_key = primary_key + pk
                END FOR

                NEW 'ClassAssertion'
                        Class_Description = class
                        NamedIndividual = primary_key
                FOREACH [ tuple as kt=>t ]
                        IF [ NOT [ kt IN array_keys(f_keys[table]) ] ]
                                NEW 'DataPropertyAssertion'
                                        DataProperty = class + '_has_' + kt
                                        NamedIndividual = primary_key
                                        Literal = t
                        ELSE
                                NEW 'ObjectPropertyAssertion'
                                        ObjectProperty = f_keys[table][kt]['pk_table']
                                        NamedIndividual = primary_key
                                        NamedIndividual = f_keys[table][kt]['pk_table'] + '_' + t
                        END IF
                END FOR
        END FOR
END FOR
//   tuples -> objectProperties (link tables)   //
FOREACH [ non_class ]
        FOREACH [ columns[table] - 1 ]
                FOREACH [ tables_data[table] as tuple ]
                NEW 'objectPropertyAssertion'
                        ObjectProperty = f_keys[table][columns[table]]['pk_table']
                        NamedIndividual = f_keys[table][next(columns[table])]['pk_table'] +
                                        '_' + tuple[f_keys[table][next(columns[table])]['pk_column']]
                        NamedIndividual = f_keys[table][columns[table]]['pk_table'] +
                                        '_' + tuple[f_keys[table][columns[table]]['pk_column']]
                END FOR
        END FOR
END FOR
```

Figure 7: Algorithm – Individuals

classes are only defined in one direction. If in the inverse pair of object properties exists one property that is functional, is that one that it is defined; if not, the generated object property assertion is irrelevant.

In the next table (Fig. 8) we summarize the mapping process. From the conceptual mapping approach and some DBMS heuristics we start to manually convert a relational database (case study database) into OWL using Protégé [40]. The algorithm was then designed based on the defined mapping and from the code analysis (Protégé – OWL/XML format).

5.2 Prototype – Tests and Results

The algorithm was then tested with the case study database. Figure 9 shows the database logical model and the ontology conceptual approach. It was necessary to do some adjustments in order to achieve a consistent ontology. Then we successfully use the HermiT 1.3.3 reasoner [41] to classify the ontology. The inverse "object properties assertions" that the algorithm do not generates for the individuals were inferred. Some equivalent (and inverse functionality) object properties were also inferred.

In Figure 10 we present an example of the generated ontol-

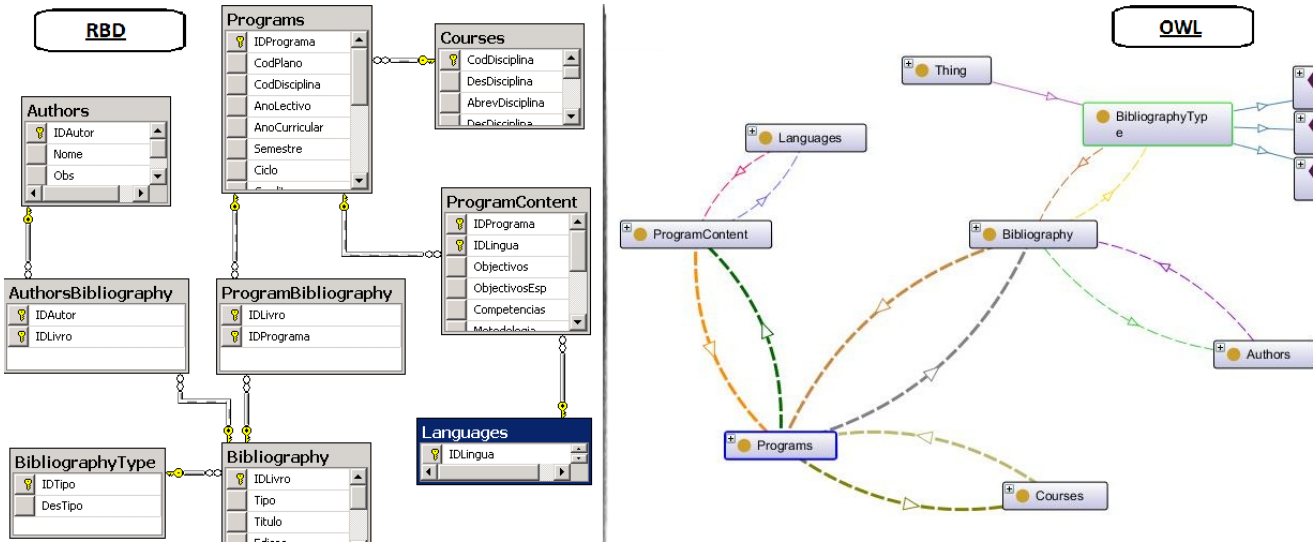


Figure 9: RDB Logical Model *vs* Ontology Overview

RDB	OWL
Tables	Classes
If (#attributes = #primary keys = #foreign keys) → Object Property	
Foreign Keys	Object Properties
Primary Keys	Individuals Identification
Other Attributes	Data Properties
Tuples	Individuals
	• Inverse Object Properties Generation • Functional Object Properties Definition • Disjoin All Classes

Figure 8: Mapping Process Sumarized

ogy. This example focus on the relationship that exists between the two tables (”Authors” and ”Bibliography”) where the link table ”AuthorsBibliography” is mapped into an object property (and inverse object property) relating the correspondent mapped classes. It is also shown a portion of the generated OWL document where we demonstrate the results of mapping a table attribute into a data property of a class.

The next step consisted on testing the algorithm with other databases. We use one MySQL database and two MSSQL Server databases (the maximum tables size were about tens of thousands records). All databases used in this research are from the University Lusíada information system.

The results were very satisfactory because the algorithm achieve similar results of the ones obtained with the case study database only with minor inconsistencies related with naming and encoding problems. The processing time is an issue directly related to the dimension of the database (it is necessary to test the algorithm with huge databases [millions of records] in machines with powerful processing capability).

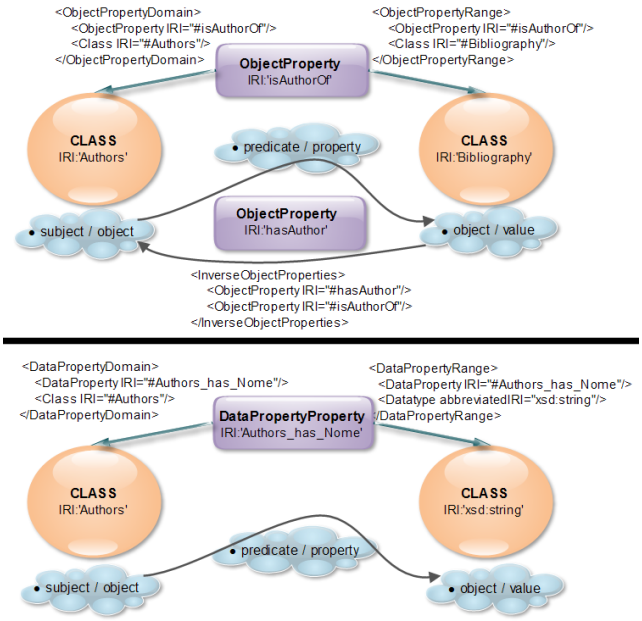


Figure 10: Results Portion: tables ”Authors” and ”Bibliography” relationship & ”Authors” attribute mapping

6. CONCLUSION AND FUTURE WORK

Ontologies and databases are related to each other because of their characteristics. Using ontologies in database preser-

vation is an approach to capture the "knowledge" associated to the conceptual model of the database.

In previous work we preserve the database data and structure (logical model) by ingesting the database in a XML based format (DBML [3]) into an OAIS [5] based archive.

Here, we present the work developed in order to convert databases to ontology, based on a mapping process (mapping algorithm), for preservation. In order to preserve a snapshot of the database (or a frozen database) we preserve the ontology (OWL [6], also a XML based format) obtained from the application of developed algorithm to the source database. We tested the algorithm with few databases and the results were acceptable in terms of consistency of the generated ontology (and comparing to the results obtained with the case study database).

This generated ontologies will induce the development of a new database browser/navigation tool.

Ontologies also have other potentialities such as the asset of providing answers to questions that other standards are limited. For example, in terms of metadata, one issue that we intend to also address in future work.

We also anticipate the possibility of integration between Web Ontology Language (OWL) and Semantic Web Rule Language (SWRL [42]) to consolidate the asserted and inferred knowledge about the database and its information system.

7. REFERENCES

[1] Pat Manson, "Digital Preservation Research: An Evolving Landscape," European Research Consortium for Informatics and Mathematics – NEWS, 2010.

[2] R. Freitas, J. Ramalho, "Relational Databases Digital Preservation," Inforum: Simpósio de Informática, Lisboa, Portugal, 2009, ISBN: 978-972-9348-18-1; [Online]. Available: http://repositorium.sdum.uminho.pt/handle/1822/9740

[3] M. Jacinto, G. Librelotto, J. Ramalho, P. Henriques, "Bidirectional Conversion between Documents and Relational Data Bases," 7th International Conference on CSCW in Design, Rio de Janeiro, Brasil, 2002.

[4] Ricardo André Pereira Freitas, "Preservação Digital de Bases de Dados Relacionais," MSc Thesis, Escola de Engenharia, Universidade do Minho, Portugal, 2008.

[5] Consultative Committee for Space Data Systems. "Reference Model for an Open Archival Information System (OAIS) – Blue Book," National Aeronautics and Space Administration, Washington, 2002.

[6] "OWL – Web Ontology Language" [Online]. Available: http://www.w3.org/TR/owl-features/

[7] A. Wilson, "Significant Properties Report," InSPECT Work Package 2.2, Draft/Version 2 (2007).

[8] Miguel Ferreira, "Introdução à preservacao digital – Conceitos, estratégias e actuais consensos," Escola de Engenharia da Universidade do Minho, Guimarães, Portugal, 2006.

[9] K. Thibodeau, "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years," presented at The State of Digital Preservation: An International Perspective, Washington D.C., 2002.

[10] "PLANETS – Preservation and Long-term Access through NETworked Services" [Online]. Available: http://www.planets-project.eu/

[11] "SIARD – Format Description," Swiss Federal Archives - SFA, 2008.

[12] XML, "Extensible Markup Language," in W3C – The World Wide Web Consortium [Online]. Available: http://www.w3.org/XML/

[13] J. Ramalho, P. Henriques, "XML and XSL - Da Teoria à Prática," FCA - Editora Informática, 2002.

[14] Michael Day, "The OAIS Reference Model," Digital Curation Centre UKOLN, University of Bath, 2006

[15] B. F. Lavoie, "The Open Archival Information System Reference Model: Introductory Guide," Digital Preservation Coalition, Dublin, USA, Technology Watch Report Watch Series Report, 2004.

[16] Tim Berners-Lee, James Hendler and Ora Lassila, "The Semantic Web", Scientific American, May 2001.

[17] Tom Gruber, "Ontology," Entry in the Encyclopedia of Database Systems, Ling Liu and M. Tamer Ozsu (Eds.), Springer-Verlag, 2008.

[18] Edgar Codd, "A Relational Model of Data for Large Shared Data Banks," in Communications of the ACM, 1970.

[19] "World Wide Web Consortium," [Online]. Available: http://www.w3.org/

[20] "Semantic Web," [Online]. Available: http://www.w3.org/standards/semanticweb/

[21] "Resource Description Framework," [Online]. Available: http://www.w3.org/RDF/

[22] G. P. Zarri, "RDF and OWL," Encyclopedia of Knowledge Management, 2006.

[23] H. Santoso, S. Hawa and Z. Abdul-Mehdia, "Ontology extraction from relational database: Concept hierarchy as background knowledge," Knowledge-Based Systems, Elsevier, 2010

[24] OpenLink Virtuoso Platform, "Automated Generation of RDF
Views over Relational Data Sources," [Online]. Available: http://docs.openlinksw.com/virtuoso/rdfrdfviewgnr.html

[25] C. Bizer, R. Cyganiak, "D2RQ – Lessons Learned," Position paper for the W3C Workshop on RDF Access to Relational Databases, Cambridge, USA, 2007.

[26] S. Auer, S. Dietzold, J. Lehmann, S. Hellmann , D. Aumueller, "Triplify – Light-Weight Linked Data Publication from Relational Databases," proceedings of WWW 2009, Madrid, Spain

[27] Farid Cerbah, "Learning highly structured semantic repositories from relational databases: the RDBToOnto tool," In Proceedings of the 5th European Semantic Web Conference, Spain, 2008.

[28] J. Barrasa, A. Gomez-Perez, "Upgrading relational legacy data to the semantic web," 15th international conference on World Wide Web Conference (WWW 2006), Edinburgh, United Kingdom, 2006.

[29] H. Chen, Z Wu, "DartGrid III: A Semantic Grid Toolkit for Data Integration," Proceedings of the First International Conference on Semantics, Knowledge, and Grid, 2005

[30] M. Fisher, M. Dean, "Automapper: Relational Database Semantic Translation using OWL and SWRL," Proceedings of the IASK International Conference E-Activity and Leading Technologies, Porto, Portugal, 2007

[31] J. Xu and W. Li, "Using Relational Database to Build OWL Ontology from XML Data Sources," CISW 2007 – Proceedings of the 2007 International Conference on Computational Intelligence and Security Workshops, IEEE Computer Society, Washington, DC, USA, (2007)

[32] G. Bumans, K.Cerans, "RDB2OWL : a Practical Approach for Transforming RDB Data into RDF/OWL," Proceedings of the 6th International Conference on Semantic Systems ISEMANTICS 10, 1-3. Retrieved from http://portal.acm.org/citation.cfm?id=1839739

[33] N. Cullot, R. Ghawi, K. Yetongnon,Ą"DB2OWL: A Tool for Automatic Database-to-Ontology Mapping," . In Proc. of 15th Italian Symposium on Advanced Database Systems (SEBD 2007), pages 491-494, Torre Canne, Italy, June 2007.

[34] "R2RML: RDB to RDF Mapping Language," W3C Working Draft, 24 March, 2011

[35] J. Trinkunas, O. Vasilecas, "Building Ontologies from Relational Databases Using Reverse Engineering Methods," International Conference on Computer Systems and Technologies Ű- CompSysTechŠ07, ACM, 2007, ISBN: 978-954-9641-50-9

[36] K. M. Albarrak , E. H. Sibley, "Translating relational & object-relational database models into OWL models," Proceedings of the 10th IEEE international conference on Information Reuse & Integration, Las Vegas, Nevada, USA, 2009

[37] C. He-ping, H. Lu, C. Bin, "Research and Implementation of ontology automatic construction based on relational database," International Conference on Computer Science and Software Engineering. IEEE Computer Society, 2008.

[38] "A Survey of Current Approaches for Mapping of Relational Databases to RDF," W3C Incubator Group, 2009

[39] I. Myroshnichenko , M. C. Murphy, "Mapping ER Schemas to OWL Ontologies," Proceedings of the 2009 IEEE International Conference on Semantic Computing, p.324-329, September 14-16, 2009

[40] http://protege.stanford.edu

[41] http://hermit-reasoner.com/

[42] "SWRL: A Semantic Web Rule Language Combining OWL and RuleML" [Online]. Available: http://www.w3.org/Submission/SWRL/

Replicating Installed Application and Information Environments onto Emulated or Virtualized Hardware

Dirk von Suchodoletz
Institute of Computer Science, Albert-Ludwigs University
10 H.-Herder st., Freiburg, 79104, Germany.
dirk.von.suchodoletz@uni-freiburg.de

Euan Cochrane
Archives New Zealand, The Department of Internal Affairs
10 Mulgrave st, Wellington, 6011, New Zealand.
euan.cochrane@dia.govt.nz

ABSTRACT

Digital objects are often more complex than their common perception as individual files or small sets of files. Standard digital preservation methods can lose important parts of digital objects, or the context of digital objects. To deal with the different types of complex digital objects, and to cope with their special requirements, we propose applying emulation from a different perspective in order to preserve the whole original environment of single digital objects or groups of digital objects. Many of today's preservation scenarios would benefit from a change in our understanding of digital objects. Our understanding should be shifted up from the single digital files or small groups of files as they are commonly conceived of, to full computer systems. When this shift in perspective is undertaken two important outcomes result: 1. the subject of preservation includes a much richer level of context and 2. the tools available for preserving them are constricted. In this paper we describe a workflow to be used for replicating installed application environments that have the x86 architecture onto emulated or virtualized hardware, we discuss the potential for automating steps in the workflow and conclude by addressing some of the possible issues with this approach.

Keywords

Emulation, Disk Imaging, Virtualization, Complex Digital Object, Original Digital Ecosystem

1. INTRODUCTION

Most digital objects are more complex than perceived by archivists or other practitioners dealing with the task of digitally preserving office workflows, scientific desktops, electronic publications or dynamic objects like multimedia encyclopedias, educational software and computer games. The majority of today's digital objects consist of individual files but most of those files are not self contained. A digital ecosystem is required to render or run these digital objects. In order to preserve the individual files, and the entirety of the information that is presented when they are rendered, it would be useful, and in many cases necessary, to be able to replicate and preserve their original rendering environments. Overall there are at least three compelling reasons for making images of entire information environments and maintaining the ability to render them over time:

1. To provide researchers the ability to experience individual users' or representative users' old information environments such as politicians', artists' and other famous peoples' information environments or an average/representative user's Information Environment from a particular time period.
2. In order to preserve complex digital objects in an inexpensive and efficient way by enabling the automation of their preservation.
3. To produce permanent "viewers" for digital objects that can easily be maintained over the long term and are known to be compatible with the objects.

Preserving a famous person's installed application and information environment has been demonstrated most successfully by the team at the University of Emory's Manuscript Archives and Rare Book Library (MARBL) where they have preserved an image of the hard disk from Salman Rushdie's early 1990s Macintosh desktop and currently use an emulator to access it [6]. The value of this approach has been realized by the team at Emory where they have seen new ways of conducting research being developed because of the availability of the emulated desktop. The experience of interacting with the original owner's actual desktop environment has led to novel discoveries being made such as the discovery the Emory team made that Rushdie was an avid user of the "stickies" application on the Mac operating system.

Unfortunately, imaging and maintaining access to old computer desktops is still a niche endeavor for a number of reasons including the perceived complexity and difficulty for average, non-technically trained preservation practitioners. This need not be the case. The steps described in this paper constitutes a feasible workflow that suitably skilled practitioners could use immediately to begin replicating installed application and information environments onto emulated or virtualized hardware. The workflow also includes numerous steps which have the potential to be highly automated, making the process easily manageable by an average archivist,

iPRES2011, Nov. 1–4, 2011, Singapore.

librarian or other less technically skilled digital preservation practitioner.

2. PRESERVING COMPLEX OBJECTS

Digital objects are often assumed to be individual computer files that stand-alone and don't have many dependencies aside from requiring some sort of program to render or "open" them.

This understanding of digital objects is very limited and quickly breaks down under closer examination. For example, there are many types of digital objects that include more than one "content files" such as digital videos which are often captured as thousands of image files with a separate audio file and separate metadata file, or linked spreadsheet-workbooks in which one table provides data to another. In examples like this the loss of any one file from the digital object can lead to the inability to render the wider object [4].

As another example, files stored in Electronic Document and Records Management Systems (EDRMS) can be ruined from an archival perspective if they are taken out of their EDRMS as their context can be lost. In this example it would theoretically be possible to define a metadata standard and preserve sufficient metadata to capture the context that the file came from; however in practice this is extremely difficult. A further example is provided by the case of office documents (e.g. Microsoft Word) that require additional/special fonts to be rendered. Without the fonts these digital objects sometimes cannot be rendered at all [1, 9]. These examples illustrate that digital objects are often more complex than they are generally conceived. In addition to this, when the necessary additional files are not saved directly with the "content files", it is often unclear and difficult to ascertain what additional components are necessary for rendering the object with integrity. This presents a real and significant problem to archivists, librarians, curators and other digital preservation practitioners who are faced with the task of preserving such objects.

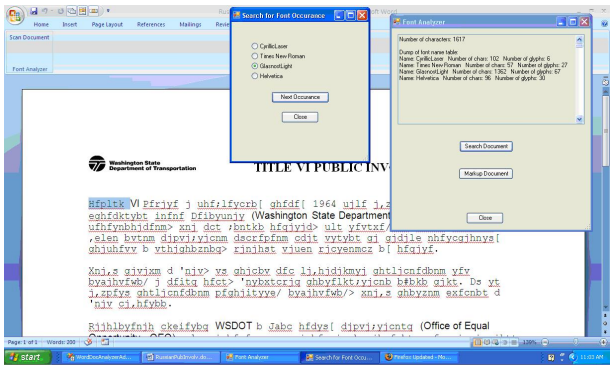


Figure 1: Missing fonts can make digital objects unintelligible [1].

All of these cases benefit from a shift of our understanding of digital objects up from the single digital files or small groups of files as they are currently conceived of, to full computer systems. By shifting our understanding up to this level we enable conversations and planning to be undertaken to identify such objects. This shift in understanding also enables preservation institutions and research organisations to begin to address the practical preservation requirements that need to be fulfilled in order to preserve such complex objects for future generations. The experience of maintaining access to old software like computer games using emulation [2, 8] helps to understand the envisioned process.

The workflow described in this paper provides one initial practical option for preserving such complex objects. The workflow can be used to preserve the complete creating and/-or rendering environment portion of a complex digital object, ensuring that any dependencies are preserved. This approach would significantly help in solving the problem outlined above.

3. PRODUCING PERMANENT, COMPATIBILITY VERIFIED VIEWERS

When addressing the problem of long term access to (or preservation of) digital objects there are two primary options that preservation practitioners have for solving it:

1. Move the information from one file or set of files to another file or set of files which can be rendered more easily using existing software.
2. Maintain the original rendering software indefinitely.

Option (1) implies many potential problems including difficulty in ascertaining whether migrated versions of objects have retained their integrity, either due to information loss in the conversion process or difficulty in assessing whether new renderers are truly compatible with the objects, and a potentially high long-term cost due to having to perform migrations on a regular (if infrequent) basis. Option (2) has been considered unfeasible for various reasons including (though not limited to) the difficulty in maintaining the viewers over time.

The workflow outlined in this paper could be used to produce viewing environments that would be known to be compatible with an organization's entire set of digital objects or digital objects across a time period. This would solve one of the problems faced by option (1) by providing a compatibility verified viewer for the objects. For example, a standard desktop environment from a public-sector organization that is intended to be able to be used to view/render all the objects created by that organization could be replicated onto virtual or emulated hardware, or a web browsing environment representing a period of time could be replicated and maintained for use indefinitely in viewing/rendering websites from that period.

Virtualized or emulated environments are designed to be portable across different types of hardware and operating systems, including potential future hardware and operating systems, and thus viewing environments created using the workflow outlined in this paper would have these same sustainability properties. Furthermore, while longevity is a general property of virtualized or emulated environments, the workflow outlined in this paper could also be used to

replicate environments for use in emulators that have been specifically designed to need very little maintenance over time, such as the Dioscuri modular emulator. This would ensure that there were even more sustainable and viable as permanent solutions. Virtualization products may be seen as being quite short-term solutions from a digital preservation perspective however they are are practical in so much as they are available now and can be used to solve current problems. Furthermore virtualization products often use emulation to provide many critical components of their total product and are best seen as emulators (e.g. VirtualBox).

4. ENVIRONMENT REPLICATION WORK FLOW

After some initial hard drive imaging and virtualization tests with a MySQL database system running on Linux to investigate the general feasibility and practicality in 2007[1] the authors undertook more elaborate and broader experiments at Archives New Zealand. In these activities we aimed to investigate the feasibility of making images of old computers that were running a wide range of DOS and Windows operating systems. The experiments aimed to replicate the installed application and information environments on the computers' disks onto emulated and virtualized hardware. This imitated the idea of moving a hard disk from one physical machine to another compatible machine, and thereby preserving most (if not all) relevant aspects of the first one. This was achievable as almost all the relevant information on a computer is kept on its hard disk. In some circumstances this approach may not be feasible due to some components being provided through external hardware or some aspects of the environment being dependant on particular firmware code, however for the most part this approach is very effective. The experiments were very successful and from these experiments a number of steps were documented and have been formed into a workflow. The workflow outlines how to replicate installed application and information environments onto emulated or virtualized hardware. A detailed discussion of the experiments, including additional findings, will be included in a forthcoming paper. The following section describes the steps involved in the workflow and outlines the options and decision points that occur within it.

4.1 Create Disk Images

The first step in the workflow is to make images of the hard drive or drives that contain the environment that you wish to replicate in emulated or virtualized hardware. There are two ways of doing this: It can be done intrusively by taking the hard drive out of the old hardware and attaching it to modern hardware, or non-intrusively, by running modern software in the memory on the old hardware and making an image of the drive(s) over a network connection.

Intrusive Disk Imaging. Intrusive disk imaging involves removing the target hard drive(s) from the original computer hardware. For this reason is both difficult for someone without specialist training to undertake and potentially quite risky as the hardware may be damaged during the process. There are two initial steps involved in this method:

1. Detach and remove the hard disk from the computer case.

2. If the connector is IDE or SATA there are then two secondary options:

   (a) Connect the hard disk to an USB-IDE/SCSI adapter and plug this one into a machine running Linux that the image is to be "dumped" (written) to.

   (b) Directly connect the drive to a machine that has a compatible IDE/SATA/SCSI connector on its motherboard or has an appropriate extension card installed and can run Linux.

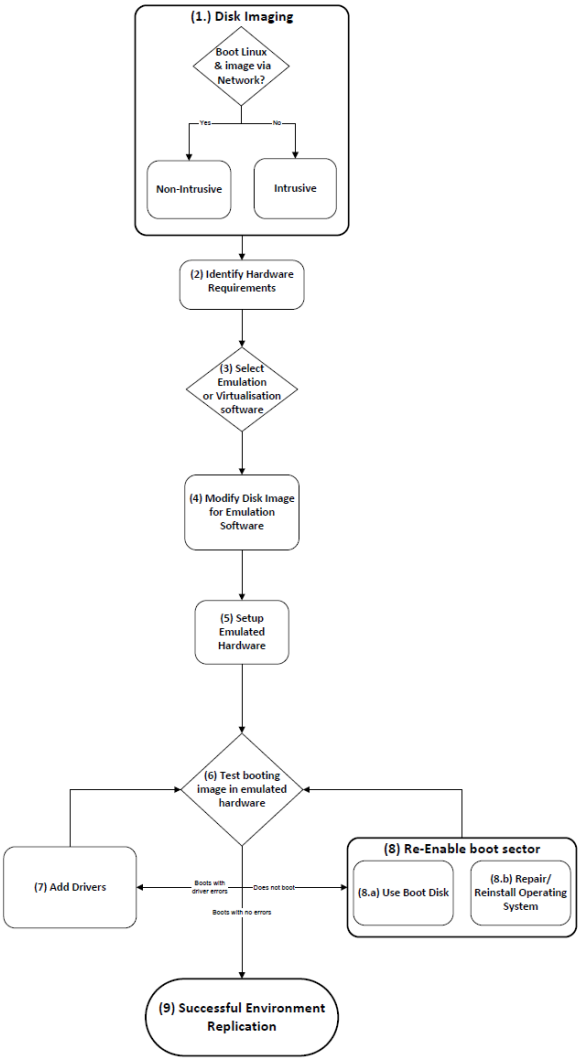


Figure 2: Formalized work flow diagram

[1] An x86 IBM server machine containing a hardware raid of three SCSI disks was dumped running in single user mode with network connection and file system set to read only. The resulting image was converted to VMware image type and the hardware driver changed to Buslogic [13, p. 165].

A computer may need to be switched off to attach or remove a disk so is a good idea to have a dedicated computer to do this with in order to avoid complications. Option 2a may be preferred in many cases as it is easier due to not having to open up the modern hardware in order to connect the old hard drive to it. However it may be substantially slower than option 2b which may be important in some contexts.

Copying the content. Once the disk is connected to a machine running Linux it is then possible to run a command to copy the entire contents and structure of the hard disk into an image file. In Linux the program that performs the imaging procedure is `dd`. To initiate the process the full command used is as follows: `dd if=/dev/sdb of=imagename.img` Where `sdb` is the logical device identifier of the attached disk in the Linux system. The disks are enumerated starting from letter a. Thus additionally attached disks are labeled in alphabetical order. Using the `fdisk -l /dev/sdb` command helps to identify the disk by its size if several are attached.

The `dd` tool reads the disk contents directly from the device blockwise from the absolute beginning to the end and thus requires administrator permissions in order to be executed. Directing `dd` to image `/dev/sdb` is the simplest imaging method as it creates a copy of the entire disk and any partitions on it. However it is possible to just make an image of the content of any relevant partitions such as imaging the second partition with `/dev/sdb2` to save on size of the resulting disk image. In the latter case the boot sector and partition table are lost though, as the disk layout is changed from a partitioned one to a linearly used disk without partitioning on it. This in turn makes running the image in an emulator much more complicated.

If `dd` encounters problems because of bad sectors then the `dd_rescue` program can be used as an alternative to help to produce an image from the readable sectors. The resulting image will be usable in many cases provided the sectors containing significant portions of the boot sector or operating system have not been affected.[2]

The time taken to image a disk using this method will vary depending on the type of connection, the size and the speed of the disk being imaged. It can take from less than one minute to more than an hour for very large disks. The process is reasonably straight forward for standard disk configurations such as those that are likely to be found in standard desktop office machines. It becomes more challenging when involving hardware or software RAID setups which might be present in servers. In these cases it might be possible to access the disks via non-intrusive imaging. Furthermore this process will likely be greatly simplified for many current day servers which are often already running from virtual disk images as virtual machines.

Non-Intrusive Disk Imaging. The least intrusive disk imaging method is to boot a "live" distribution of the Linux operating system into the memory of the computer, and image the hard disk over a network connection directly to another machine (Fig. 3). These "live" distributions do not use the internal hard disk and thus preserve its original state and leave it free (not in use) to be imaged. The "live" distributions come in a number of forms that can be run on varied sets of hardware. Linux includes a wide range of popular hardware drivers. Support for standard IDE and SCSI disks will be included in most distributions; many hardware RAID controllers are also supported. Most versions of "live" Linux distributions consist of a CD-ROM which can be booted from. It is also possible to find distributions that include a boot-floppy disk for machines that require this to be run first in order to boot from the CD-ROM. For machines that can boot from USB, Linux distributions that run from USB drives can be found. If the machine is an older one without a CD-ROM drive, a compact Linux distribution can be used that will run from one or more floppy disks such as MuLinux.[3]

Figure 3: Booting Damn Small Linux on an x86 Compaq desktop machine to allow non-intrusive disk dumping

Once an appropriate Linux distribution has been selected it needs to be booted in "live" mode on the original hardware to a point where the command line is accessible. The hardware needs to be connected to a *dumping* target machine via a network connection and the connection has to be established between the two.[4] The Linux commands `ifconfig` or `ip` can be used to setup the connection, assign an IP address to the hardware containing the target hard disk. The connection can be tested using the `ping` command. If there is no Linux hardware support available for a given disk configuration it

[2]The `dd_rescue` program, unlike the `dd` program, is not normally included by default in Linux distributions.

[3]See: *http://www.micheleandreoli.it/mulinux*

[4]We keep a selection of Linux supported NICs: PCI, Cardbus, ISA to install them into the target if required.

Tool / Image type	dd-raw	vmdk	vdi
QEMU	x	x	x
VirtualBox	–	x	x
VMware	–	x	–

Table 1: Compatibility list of container formats understood by the different x86 virtualization tools or emulators.

would be possible to find other disk imaging tools which produce similar results.[5]

Copying the Content. As when imaging a hard drive that is directly connected to modern hardware, the command to use on the just started Linux is `dd`. To initiate the process the full command used for the machines that is to send the image, via SSH (a secure data exchange protocol), to the image storage computer over the network will be similar to this:

```
dd if=/dev/hda|ssh username@ip.of.target.machine
dd of=imagename.img
```

Where the *username* was the username to be given for the remote machine, the *ip.address.of.target.machine* is the IP address of the computer the image data is being sent to, and the *imagename.img* is the name that was to be given to the file to that the image was to be written to. Some older Linux kernels might use different naming for IDE disks like *hda* instead of *sda*. `fdisk -l` usually gives the information on the installed disks of a computer.

The time taken for this process will vary significantly depending on the throughput of the network connection and the size of the hard disk being imaged, from minutes to hours.

4.2 Emulation Software Preparation

Once a copy of the original hard disk has been written to an image file the steps to resurrect the environment in emulated or virtualized hardware can be begun. The first step in this sub-process is to identify the hardware requirements of the software environment that is being replicated. In order to select an appropriate emulation or virtualization application it is first necessary to identify the hardware requirements of the software environment to be replicated so that they can be correlated to those provided by the different emulation and virtualization applications. The choice of software to be used to provide the virtualized or emulated hardware to replicate the imaged application and information environment on will depend on a number of factors.

Hardware Provision. Availability of necessary virtual or emulated hardware is one of the primary considerations when selecting a virtualization or emulation application for long term preservation/access purposes. For example, if the environment being replicated is to be used to access networked resources such as old websites, then it will be necessary for the emulation or virtualization application to provide a virtual or emulated network card that is compatible with the operating system of the environment that is being replicated. Depending on the virtual/emulated machine, numerous hardware configuration options are available:

- QEMU[6] supports numerous different graphic cards like Cirrus gl5446, generic VESA or the VMware SVGA II and different sound and network adapters (ne2000, AMD PCnet, rtl8029, rtl8139, Intel e1000, ...)
- Virtual PC offers S3 Trio32/64 graphics adapter
- VMware Workstation provides an SVGA adapter without a real world counterpart and different kind of network adapters: the AMD PCnet, the Intel e1000 and virtual IO.
- Virtual Box is flexible regarding the chip set (PIIX3,4 and ICH) and other options like Soundblaster 16 and ICH AC97 audio.
- DosBox[7] implements S3, et3000, et4000 and other graphics adapters, various sound cards and a higher layer networking.

Operating Systems and Hardware Drivers. Hardware and operating systems were once tightly matched for a number of older computer platforms such Motorola or PowerPC0-based Macintoshes or Atari home computers. This was not the case the with x86 architecture ("PCs"). The system BIOS of modern computers provides a number of standard APIs for the operating system to access the hard disk, floppy drive and graphic adapter. The operating system only needed specific hardware drivers to exploit the full feature set of other components like network and audio adapters or a graphic card's 3D capabilities.

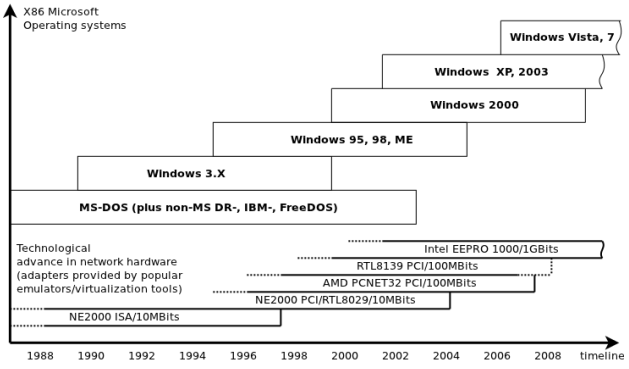


Figure 4: Driver support in standard x86 operating systems compared to provided emulated network adapters

[5]Open Source solutions like Clonezilla, *http://clonezilla.org* or commercial products Norton or Symantec Ghost and alike

[6]A popular Open Source multi-platform emulator for X86, PowerPC, ARM, S390 and others, see *http://wiki.qemu.org*. See for DP considerations [14].

[7]Open Source x86 and DOS emulator for BeOS, Linux, Mac OS X, OS/2, and Windows, *http://www.dosbox.com*.

Unfortunately, the driver situation of commercial operating systems for older x86 architecture computers is rather diverse [13, p. 190-195]. There is no driver support in DOS or early Windows versions for modern hardware. The emulated or virtualized hardware set must match the era the operating system was available and so emulators have to provide a wide range of different hardware configurations if they are to be able to host the x86 operating systems covering more then two decades. (Fig. 4).

Tool Considerations. Another major consideration is longevity. Software that provides emulated hardware is the only viable option for digital preservation as emulating hardware does not require running the emulator application on hardware that is compatible with the emulated hardware. Virtualization, on the other hand, relies (for the most part) on the underlying hardware, which the software that provides the virtualized hardware runs on, to be compatible with the virtualized hardware (e.g. to run an application that provides virtualized PowerPC hardware the underlying hardware on which the application is run must also be PowerPC compatible). A further consideration is licensing cost: Those vary amongst virtualization and emulation vendors and over their numerous products and services, and in some cases this may restrict available options. However in other cases it may be advantageous for institutions that already use particular virtualization software to continue to use this software for their digital preservation needs over the medium term (over the long term emulation is the only option as identified above).

Modify Disk Image Format for Emulation or Virtualization Software. The QEMU tool suite provides a tool "qemu-img" to handle disk image files from a wide range of virtualization tools and of course QEMU itself. The terminal command used to convert a raw image file using qemu-img would look similar to this depending on the expected outcome: `qemu-img convert -O qcow2 nameOfImageFileToConvert.img nameOfConvertedImageFile.qcow2`.

This produces a converted image file in the most recent QEMU disk container format. It can also be given a parameter that turns on compression to reduce the file space taken. For converting to formats compatible with VMware and VirtualBox, `qemu-img convert -O vmdk nameOfImageFileToConvert.img nameOfConvertedImageFile.vmdk` produces disk images usable by the VMware virtualization tool suite or by Virtual Box. Conversion directly to the native format of Virtual Box and Virtual PC is also possible, however in tests run in the laboratory VirtualBox often failed when emulating older operating systems and Virtual PC is mostly deprecated. For all experiments involving the alteration of the hard disk image file it is a good idea to also keep the original. Some experiments such as starting QEMU directly on a container file to check if the installed operating system boots, may render the image unusable for further experiments.

After producing the proper container format for the particular virtual machine or emulator the next step is to configure the virtual or emulated machine to boot from the image. The steps involved in this process depend heavily on the tool in use: While QEMU is normally configured solely via a command line most of the tools require a setup procedure using some sort of Graphical User Interface (GUI) and these are available for QEMU also). The virtual hardware configuration should match to the capabilities of the original operating system and it's supported hardware components. For example it is easier to adapt a Windows 95 based system to the emulated PCnet network adapter than the Intel e1000 network adapter as the PCnet adapter is of the same era as Windows 95 and has driver support included with the operating system.

4.3 Test Booting of Images

Once the emulator or virtual machine is properly configured, and the disk image file correctly linked to it, an attempt can then be made to boot the original system in the emulator or virtualization software. If the virtual environment boots successfully with no errors then pending further tests the process has been successful and is now complete. There is a possibility that the environment may simply fail to boot from the beginning. This can often be caused due to changed disk geometries or different BIOS capabilities of the original and the virtual machines. When this occurs the boot sector may need to be re-enabled.

The system can often be made bootable by using a valid operating system boot medium such as a Windows boot floppy disk or, for newer systems, the optical installation medium (e.g. the Windows CD-ROM). Another option when these issues are encountered is to create a system boot disk on the original system that was imaged. This has the advantage of matching exactly with the installed software. Alternatively disk images of boot-media for most operating systems can be download from the Internet. In the experiments that

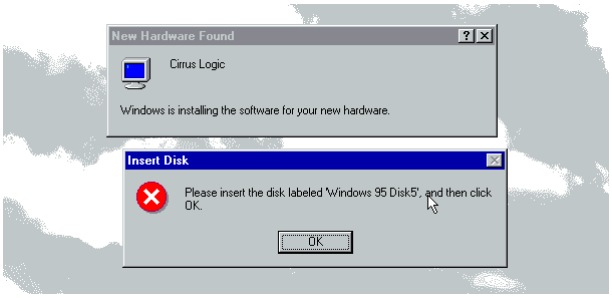


Figure 5: Reinstallation of drivers on emulated/virtualized hardware

were conducted, almost all of the system images were able to be booted directly within a QEMU-emulated machine and the other virtual machines tested. Only the Windows 98 system required a boot floppy disk to be used to restore the boot loader. This was achieved by loading the image with the boot disk in the emulated floppy drive, and typing: `a:` followed by `sys c:`. This reinstalled the necessary portion of the operating system core files and made the image bootable again. These actions might differ for other operating systems and could be more complex for newer systems like Windows XP, OS/2 or Linux. The next issue that may arise is an incompatibility between the installed hardware drivers and the new hardware provided by the emulation or virtualization software.

4.4 Finalize the Migration

Typically the emulator or virtual machine being used will get to a point at which it produces errors about missing hardware, or wrong drivers or will not boot into the original system's GUI at all. In order to rectify this, the next step required is to re-run the operating system's hardware setup procedure to re-detect the hardware configuration (Windows 95 and above). This usually triggers the reinstallation of drivers (Fig. 5) which in turn should give improved VGA, sound output and network access. As the drivers are often not part of original installation these need to be provided either by preinstalling them on the disk image or by making them available via the virtual/emulated CD or floppy drives.

Other systems such as Windows 3.X can require the Windows setup procedure to be run before loading the operating system in order to swap the video driver back to a basic VGA driver (Fig. 7, otherwise it may be impossible to load the OS in a meaningful way). If a higher resolution and/or color depth is desired this can then be achieved through changing the settings via the GUI once VGA mode has been enabled (Fig. 6). Unfortunately, successfully making the image bootable does not always mean the operating system will be able to completely load from the image. Hardware compatibility errors will often arise from attempting to run an operating system configured for one set of hardware on a different set of hardware. Typically the emulated machine being used will get to a point at which it produce errors about missing hardware, wrong drivers or will not boot into the original system's Graphical User Interface (GUI) at all. The solution to this (for Windows 95 and above) is to run the operating system's hardware setup procedure to re-detect the hardware configuration. This usually triggers the re-installation of drivers which in turn should give improved VGA, sound output and network access. As the drivers are often not part of original installation these need to be provided either by preinstalling them on the disk image or by making them available via the virtual or emulated CD or floppy drives.

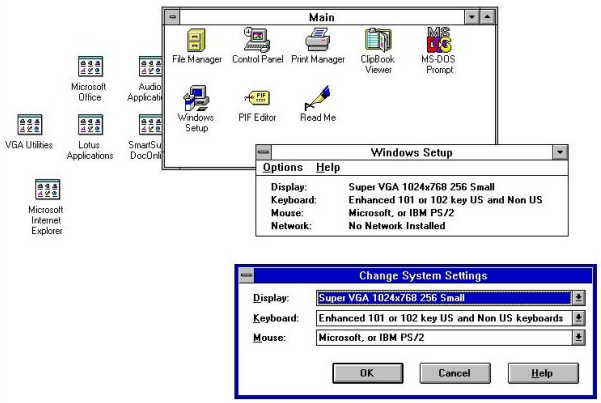


Figure 6: Alterations to Operating System Configuration

For newer proprietary operating systems, the original installation media or some equivalent is usually require to provide the standard driver set. Additionally, if the hardware changes the license might needed to reactivated. This requires the license key to be available to be re-entered. In order to make this step easier it is often possible to find software tools to read license keys read from the hard drive image from where they are stored on the disk, for example from the Windows system registry.

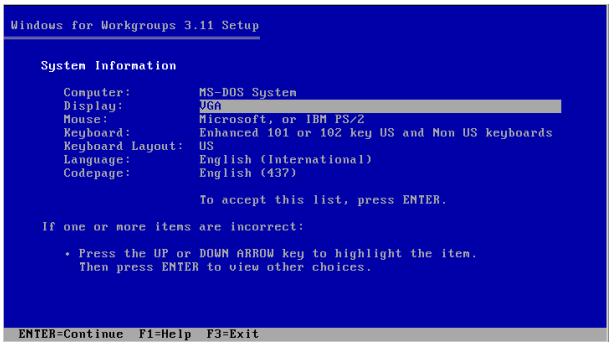


Figure 7: Altering Display Drivers in Windows 3.11

Fortunately, there is not a huge degree of variance between most hardware components in the different emulated hardware environments. Because of this, it should be possible to create additional disk image containers for each emulator or virtualization tool containing all of the relevant files that are needed to revive an arbitrary imaged original environment. Typical files which should be included are the installation files, relevant drivers and additional utilities which might help later users of the imaged systems. In most cases these files could be copied to a blank formatted hard disk image and such a hard disk could be attached as a further device to the virtual/emulated machine which would be automatically recognized by most, if not all, operating systems. At this point the environment should be ready for use.

5. POTENTIAL FOR AUTOMATION

The research conducted in order to inform the creation of the workflow outlined in this paper also enabled the testing of the general feasibility of dumping complete machines to be run in emulated hardware environments. One result of these investigations was the realisation that some of the tasks may be seen to be somewhat complicated for a non-technically trained practitioner (such as the use of the command line). Another finding was that there are many steps in the workflow outlined above that have the potential to be fully or partially automated, often by rather simple batch scripts run within an average Linux environment. Future research would be required to establish the best techniques for handling a wider range of systems and for dealing with more variable hardware. It would be beneficial to allocate resources to investigate how to automate some of the steps within the original environments. Here it might worthwhile to investigate whether the results of previous research into handling of interactive environments could be applied for automating some of these steps.

Step 1(.2) (Non-Intrusive) Disk Imaging could be automated through the creation of customized Linux distributions that are setup solely for the purpose of imaging old media. Linux distributions designed for digital forensics workflows [5] could

be modified for this purpose. This part of the process could be highly automated to a point where it would simply be a matter of inserting the disk or CD into the target computer, connecting a network cable and typing a few short commands, or making a few selections, to adjust the settings for the particular situation.

Step 2 Identify Hardware Requirements could be automated through the creation of an application to scan the disk image to identify most or all of the hardware requirements of the software environment installed within it. For example Windows operating systems often store information about the hardware that the operating system is installed on within the system registry files and these files could be queried to provide information on the appropriate emulation environment. If the machine is accessed via Linux for dumping, than listing of hardware components by using standard hardware detection tools like `hwinfo` or `lspci` is pretty straight forward.

Step 3 Select Emulation or Virtualization Software will always be a decision point, however intelligence could be added to make the decision as simple as possible by having the results of step 3 automatically compared to a database of virtual and emulated hardware provided by the different applications. Extended tool registries like [10] or [11] should be able to support this in the future.

Step 4 Modify Disk Image Format for Emulation or Virtualization Software could also be automated through the creation of an application that took the results of step 3 and automatically converted the disk image to the appropriate format. As this is run on a today's machine it is just a line to be added to a script running on the emulator hosting computer.

Step 5 Setup Emulated or Virtualized Hardware could be partially automated by mapping the results of the hardware identification in step 2 to the settings in the virtualization or emulation software being used. There will still likely need to be some decisions made about the configuration for each different environment but where generic enough this may be able to be fully automated.

Step 6 Test Boot of Image in Emulated or Virtualized Hardware has the potential for automation through a number of mechanisms. The error logs of the virtualization or emulation software could be analyzed to check for any boot-problems. It may also be possible to identify drivers by analyzing the disk image file to ascertain hardware conflicts in the installed operating system after an initial boot test, or by applying image analysis using the I/O interface (such as VNC or RDP) of the emulation or virtualization software to check for errors being presented in the virtual or emulated environment.

Step 7 Add Drivers could be further automated by the creation of custom applications for each operating system and virtual/emulated hardware combination that could be run on the virtualized/emulated environment to install the necessary applications. Media containing the necessary drivers could be attached to the virtual/emulated system and the relevant options could be automatically selected when the operating system asks for the necessary driver files.

Step 8.1 Re-Enable Boot Sector could be automated using the I/O interface of the emulation or virtualization software. This could be accomplished by the use of methods similar to those identified in previous papers published by one of the authors on using emulation for bulk migration [13]. This method uses an program running outside of the emulated or virtualized environment to automatically interact with the environment and select the relevant options when necessary during the boot-sector re-enabling process.

6. OTHER CONSIDERATIONS

Besides the technical procedures a number of additional issues should be considered by memory institutions dealing with emulation of original environments.

Technical Expertise. As in other domains of non-digital and digital preservation, specific expert knowledge is required to execute the workflow outlined above. This includes a basic understanding of computer architectures like x86 or the different Apple Macintosh platforms. Skills for disassembling old desktop or rack mount computers or laptops of various kind are needed to handle the hardware and in particular the hard disks with the required care. In order to successfully execute this workflow, future digital archivists and archaeologists will need to have a good knowledge of the operating systems of the past, at least to the degree that they are able to identify the vital parts to make them executable again on the emulated or virtualized hardware. The authors were able to handle the aforementioned DOS and Windows environments. Nevertheless dealing with OS/2 and BeOS (two other operating systems popular on the x86 platform mid to end of the 1990s) was placed out of scope of the tests as specific expert knowledge was not available. While the x86 and Motorola CPU based Apple platforms were reasonably mainstream, the preservation of other architectures like Sun Solaris on Sparc or DEC Alpha would require experts both of the platform itself and of the emulators involved.

Legal Issues. Beside the technical challenges a range of legal implications need to be considered. While in theory the moving of a software installation from one computer to another without duplicating it was not prohibited by the original license terms of many older software applications and operating systems, the whole domain is fairly undefined as yet [12]. Newer licensing terms are more restrictive and often require a re-licensing action when changing hardware. Only the license agreements of newer operating systems tend to explicitly deal with the option of virtualization. Furthermore, the preservation of entire original environments requires additional workflows beside the technical ones described in this paper. Not only does the hard disk need to be copied, but all the licenses of the installed software components need to be transferred to the receiving memory institution. This shouldn't be a problem in most cases, as the software is typically deprecated and not used anymore by the donor or transferring agency. But it implies or requires additional procedures to cover such things as the transferring over of license material and software keys. Those items need to be stored with the metadata of the original environ-

ment in order to best ensure their long-term preservation. The possibility and implications of running several instances of the same machine need to be considered also as this, while potentially very useful, will likely be illegal in most jurisdictions without the purchase of additional license keys.

A completely different legal issue exists regarding the privacy concerns of the donors. There are unlikely to be many privacy issues for government archives taking transfers of official government equipment, as users are and were typically prohibited from using their machines for private matters. Unfortunately it is a completely different situation regarding computers donated by famous authors or politicians. As the system imaging uses well established procedures of computer forensics the contents of resulting image file is often complete in a way that is beyond the imagination of the original user [7, 5]. Depending on the file-system and applications originally used, a great deal of additional information can be included than might be expected by a donor. Files are not deleted instantly in many file-systems but blocks containing them simply marked as empty. They don't get overwritten until the file space is required. Thus many deleted or other types of temporary files can easily be restored from the disk image files by experts. Thus special routines might be required to "clean" the resulting file image.

Authenticity Challenges. Part of the intent of the procedure suggested in this paper is to ensure the ability to preserve access to versions of digital objects that can be verified to have maintained their significant properties and thus verified to be authentic. [3]. In general there is a greater likelihood that digital objects preserved using this method will have full information integrity due to the objects being presented to users using their original software environments. This outcome is challenged to a varying degree in three ways:

- The original system can be altered by the re-installation of necessary hardware drivers and required adaptations to the new virtual hardware environment.
- Privacy concerns might require the removal of sets of files or ultimately, the cleaning of certain file-system blocks.
- Legal restrictions may require the removal of once installed applications or software components.

These threats to the authenticity and integrity of the environments, and the objects that are viewed and interacted with using them, are important but are not devastating to the outcome. In most cases the changes that need to be made to the driver files, or system set-up files, will not cause any difference to be manifested in the final performance of the replicated environment. Furthermore, in the cases where the aim of the system replication is to produce a representative desktop environment from an organization, it can be argued that such environments ought only pass the same tests of compatibility with the new hardware as the original environments did. In such cases it may be sufficient to confirm that the operating system is adequately running on the new emulated hardware as that would have been the only test any particular PC had to pass in the organization from which the replicated environment was representing an example. Usually most of those adaptations do not affect the rendering experience of objects. Nevertheless lower or higher screen resolutions, different color depth or the availability of 3D rendering may alter the overall experience of certain object classes. One counter argument to this is provided by the fact that in most organisations that created older digital objects, any particular user was only expected to have a representative rendering environment for the objects that they dealt with. For example in a most organisations there would have been many different hardware environments being used to render the same digital objects through the use of sharing technologies such as floppy drives and/or network connections. For this reason it can be argued that preservation practitioners should not have to fulfil any greater requirements when undertaking the preservation of the objects created in such environments. In other words, it can be argued that preservation practitioners only need to provide a representative rendering environment as that is all that an average user from the time of the creation of the objects was expected to have.

For the other two challenges, the redaction procedure could be built in to transfer/donor agreements such that memory institutions had the donor or transferring agencies approve any data redaction or software redaction that had to take place. Furthermore challenge three may well end up not being an issue for some institutions if laws can be changed to enable emulation to take place without the burden of license payments.

7. CONCLUSION

By demonstrating the feasibility of x86 system imaging and reproducing the imaged information environments (Fig. 8) in emulated hardware environments, the authors demonstrated an alternative understanding of complex objects that required a specific type of preservation solution. The focus is shifted from the characterisation of objects in attempts to make them reproducible in completely different digital ecosystems, to the preservation of the whole original environment in which the objects were created, managed or viewed. Using this new technique no specific knowledge of the object and creating application is required. As emulation and virtualization of the x86 architecture is well established, the described method might be used to simplify certain preservation workflows. By utilizing this approach the current diverse methods for the handling of different types of digital object have the potential to be simplified into a standard procedure for preserving a whole computer.

Preserving significant properties and object experience using emulation is discussed in [2], [8] or [9]. Depending on the object it may be desirable to alter the emulated hardware configuration to its needs instead of adapting the original environment to the hardware set provided by the emulator. This would be achieved by writing additional components for emulators so that they emulate the specific hardware of the environment that has been imaged. This is definitely a more costly option: if the demand for emulation increases more funding will be sought and found to provide just such solutions. It is also true that there is potential to share the costs in any such endeavour so as to get great benefits for all involved for little relative cost to any particular contributor.

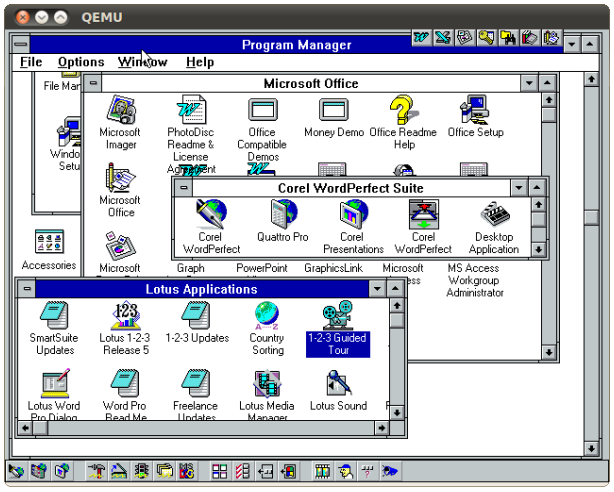


Figure 8: Imaged mid-1990ies Compaq desktop (Fig. 3) Windows 3.11 system running in QEMU

There is little difference between preserving a representative Windows 3.11 PC desktop from a government agency and preserving the system images of the portable computers of famous authors or important politicians. The workflow outlined in this paper can be applied to both situations.

The number of operating system and computer architecture combinations is much smaller than the number of file formats or complex digital objects like computer games to be considered. And, the emulation expertise could be easily shared between memory institutions. The workflow for replicating installed application and information environments onto emulated or virtualized hardware that is outlined in this paper should be able to be immediately tested and integrated into the digital preservation business processes of organizations where the necessary expertise and equipment already exist. For those where the workflow appears daunting or difficult it may be possible to obtain staff with the necessary expertise either temporarily or permanently if this is likely to be a regular process. For those where expertise is not likely to be available for some time, it should still be considered as an option when acquiring digital objects as there is significant potential to automate much of this workflow such that the expertise required will be minimal at a future date.

Automation of the aspects of the workflow that are identified above should be a primary research and development objective for the digital preservation community. The options and cost savings that the availability of replicated installed application and information environments provide, though not discussed in detail in this paper, are too large to be neglected. In spite of all the benefits of the approach outlined in this paper, a weak point still exists in the inability to be sure of the permanent availability of suitable emulators. While the number of short-term solutions in the virtualization sphere is quite high, a long-term, comprehensive digital preservation-aware emulator is still to be created. However in the meantime the number of good open source emulators could very well bridge the gap or grow into sustainable solutions [14].

8. REFERENCES

[1] Geoffrey Brown and Kam Woods. Born broken: Fonts and information loss in legacy digital documents. *International Journal of Digital Curation*, 6(1), 2011.

[2] Mark Guttenbrunner, Christoph Becker, and Andreas Rauber. Keeping the game alive: Evaluating strategies for the preservation of console video games. *International Journal of Digital Curation*, 5(1), 2010.

[3] Helen Hockx-Yu and Gareth Knight. Automation of flexible migration workflows. *International Journal of Digital Curation*, 3(1), 2008.

[4] Aaron Hsu and Geoffrey Brown. Dependency analysis of legacy digital materials to support emulation based preservation. *International Journal of Digital Curation*, 6(1), 2011.

[5] 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.

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

[7] Sumit Paul-Choudhury. Digital legacy: Respecting the digital dead. *New Scientist Online*, 2011.

[8] Dan Pinchbeck, David Anderson, Janet Delve, Getaneh Alemu, Antonio Ciuffreda, and Andreas Lange. Emulation as a strategy for the preservation of games: the keep project. In *DiGRA 2009 – Breaking New Ground: Innovation in Games, Play, Practice and Theory*, 2009.

[9] Thomas Reichherzer and Geoffrey 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.

[10] The National Archives TNA. The technical registry pronom. Online, *http://www.nationalarchives.gov.uk/pronom*, 2010.

[11] UDFR Interim Governing Body. Unified Digital Format Registry – UDFR. Online, *http://www.udfr.org*, 2011.

[12] Jeffrey van der Hoeven, Sophie Sepetjan, and Marcus Dindorf. Legal aspects of emulation. In Andreas Rauber, Max Kaiser, Rebecca Guenther, and Panos Constantopoulos, editors, *7th International Conference on Preservation of Digital Objects (iPRES2010) September 19 - 24, 2010, Vienna, Austria*, volume 262, pages 113–120. Austrian Computer Society, 2010.

[13] Dirk von Suchodoletz. *Funktionale Langzeitarchivierung digitaler Objekte – Erfolgsbedingungen für den Einsatz von Emulationsstrategien*. Cuvillier Verlag Göttingen, 2009.

[14] Dirk von Suchodoletz, Klaus Rechert, and Achille Nana Tchayep. QEMU – A Crucial Building Block in Digital Preservation Strategies. In Wolfgang Müller and Frederic Pétrot, editors, *1st International QEMU Users' Forum – DATE 2011 Workshop*, Grenoble, France, 2011.

Remote Emulation for Migration Services in a Distributed Preservation Framework

Dirk von Suchodoletz
Department of Computer Science
University of Freiburg
Freiburg i.B., Germany
dirk.von.suchodoletz@rz.uni-freiburg.de

Klaus Rechert
Department of Computer Science
University of Freiburg
Freiburg i. B., Germany
klaus.rechert@rz.uni-freiburg.de

Isgandar Valizada
Department of Computer Science
University of Freiburg
Freiburg i. B., Germany
isgandar.valizada@rz.uni-freiburg.de

ABSTRACT

Previous studies have shown the feasibility of migration services when using emulation technology. To make rather complex setups of original digital ecosystems usable in standard mass migration workflows, a separation of the systems running and their interfaces is required. Here remote emulation can become a crucial building block of future distributed preservation workflows and access systems.
In this paper we develop the requirements and a component model for *migration-by-emulation* services in a distributed environment based on division of labour. We suggest a modular system which offers interfaces to be accessed by standard preservation frameworks. It provides Web services allow access to original system environments via emulation engines with additional methods for automated interaction. The proposed migration units support versatile migration services and offer a wide range of file conversions based on a digital artifacts' original applications. The component-based architecture allows the distribution of system components among specialized memory institutions.

Categories and Subject Descriptors

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

Keywords

Digital Preservation, Emulation, Migration, Workflow, Automation, Original Environments, System Design

1. INTRODUCTION

Access to digital objects (DO) is at risk if their format is no longer supported by actual digital environments. Preserving digital artifacts for future access requires long-term support for applications and/or operating systems capable of accessing or running them. The emulation of original hardware/software environments provides the opportunity of using objects in their creation environment.

In most cases the applications or operating systems developed by the format vendors or software producers are the best candidates for handling a specific object of a certain type. The vendors are expected to have the most complete knowledge about their own data formats and the information available publicly is often incomplete or non-existent, especially regarding proprietary formats. Thus, in many cases there are no alternatives to access those objects within their original environments.

However, the ability to access obsolete DOs only by their corresponding original applications limits future access, especially if a certain environment is not available any more. This hindrance to future usage could be overcome by either migrating the object into an actual format accessible with todays tools or using its original environment. Emulation is the best way to reproduce original environments, which themselves provide the base layer for very flexible multiple migration input-to-output format scenarios. Furthermore, such emulation-based migration paths can be verified and evaluated in terms of quality and costs like traditional command-line conversion tools.

Performing migrations manually for every digital object is not a feasible strategy in many cases. Due to the large quantity of DOs held by many institutions, it would become a time-consuming and costly task. Additionally, depending on the original environments, many archivists or private users aren not knowledgeable about how to install a certain application or operating system or how to handle a certain emulator. Thus, a major prerequisite for making use of the emulation of original digital ecosystems in digital preservation (DP) workflows is the separation of the system running from the user in- and output. This allows the offering of emulation services over the network and the automation of user interaction. Based on previous studies, a system framework is required, such that complex migration tasks can be carried out in a scalable and controllable way. It should be possible to plug these services into existing preservation frameworks such as provided by PLANETS [8] or Rosetta.[1]

[1]A preservation framework developed and marketed by ex Libris.

iPRES2011, Nov. 1–4, 2011, Singapore.

Following the aforementioned discussion, the suggested component system should meet a number of requirements:

- *Large Scale and Large Quantities*
  In order to provide large-scale migration facilities based on emulation, atomic migration components are required. Each atomic migration unit could be be run in parallel, and therefore being applicable for Cloud scenarios or the like.
- *Large Variety of Migration Paths*
  By utilizing a distributed framework, more complex migration paths become possible. Institutions can specialize in special subsets of digital objects or system environments. These can be made accessible as atomic migration units, which then can be connected to complex migration paths and DP workflows in general.
- *Accessibility for a Wide Range of User Groups*
  Through decentralized and specialized competence centers in emulation, practical knowledge is made available on the ancient system environment preparation, running, and maintenance, and applications can be shared within the DP community. Migration and verification workflows based on emulation thereby become more (cost-)efficient, especially for small tasks.

2. RELATED WORK

Remote emulation has been a topic in digital preservation for a while. Several approaches have been explored and some stable prototypes have been shown to be quite feasible. There are several ways to achieve the separation between a server able to run the complex tasks and a rather simple client mainly limited to user in- and output. This concept of separation is used more and more often to run complex computer games over the network without requiring the installation and permanent updating of client software on the user's side. Several ways of implementing a remote emulation service have been researched in the last couple of years. At first this research focused on direct user interaction with the original environments [18, 9, 7]. More recently, other uses like automation of workflows have become relevant [23].

The variants of a remote emulation service or emulation streaming service can be distinguished in several ways. There are various approaches to the transport protocol used to send the screen output to the user's device and receive the user's input. Some of the protocols are public and many open source implementations exist [24]. Others may have the added functionalities of audio stream transportation, removable block device data, or implement remote USB.

At Victoria University in Wellington, New Zealand, a prototype was implemented to demonstrate the feasibility of running a game on the visitor's smartphone without the need to install any software besides a simple streaming client [4]. A very popular method for remote access is the VNC protocol [13] implemented in a wide range of operating systems and appliances. Some emulators like QEMU or virtual machines like Virtual Box and VMware implement direct access to the virtual screen and input devices. Additionally, VNC was added to Dioscuri [6] as an outcome of a student's thesis. Since VNC implements mostly screen rendering plus mouse and keyboard input, other remote protocols like RDP or Citrix[2] seem to be more attractive as they are capable of transporting audio streams or even block devices over the net. But they are proprietary and the number of server and client implementations is comparatively limited. Since they are dependent on the direct implementation of remote access in the emulator, a large number of them can not be used. Following the research into GRATE [24] during PLANETS [5] a VNC enabled prototype of a remote emulation was developed for the OPF[3]: It realizes two types of services – a migration-by-emulation service and a create-view service (e.g. [11]).

3. DISTRIBUTED MIGRATION BY EMULATION

Migration-by-emulation describes the concept of using the original or a compatible environment of a designated digital object running in a virtual machine and thus replacing the original hardware and/or software stack. This approach avoids the often impossible alteration and adaptation of outdated software to present-day environments. An abstract and generalized migration-by-emulation workflow is depicted in Figure 1. A virtual machine runs within the host environment, which contains the selected original system environment suitable for handling a certain type of digital objects. The original system environment is either reproduced from original software stored in the software archive or cloned from a prototypical original system (cf. [19, p. 165]). The selection of an appropriate system environment for each object type can be described as a so-called *view path*, a pathway pointing from the digital artifact into its original rendering or execution environment [17].

To make migration-by-emulation deployable in large-scale preservation scenarios without relying on user interaction, the user's function is replaced by a workflow execution engine [12]. This requires appropriate interfaces in order to use emulators [20]. In contrast to simple command-line input-output migration tools, a migration-by-emulation service needs a more complex initial setup:

- *System Emulation*
  Hardware emulation including a full reconstruction of outdated components. For instance, a i386 CPU, ISA Systembus, PS/2 mouse and AT keyboard and VESA compatible graphics are minimal requirements, e.g. for Windows 3.11.
- *System Environment*
  An appropriate runtime environment (e.g. a disk image file) preconfigured with operating system, necessary drivers and tools, and the required target application. Furthermore, each environment specifies at least one transportation option, defining how digital objects can be injected into and extracted from the virtual environment. Examples range from different kinds of floppy-disk images to hard disk container formats and advanced networking options.

[2]cf. [24] for an overview and evaluation of the various protocols and their usability in remote emulation.
[3]Open Planets Foundation, non-for profit PLANETS follow-up, *http://www.openplanetsfoundation.org*

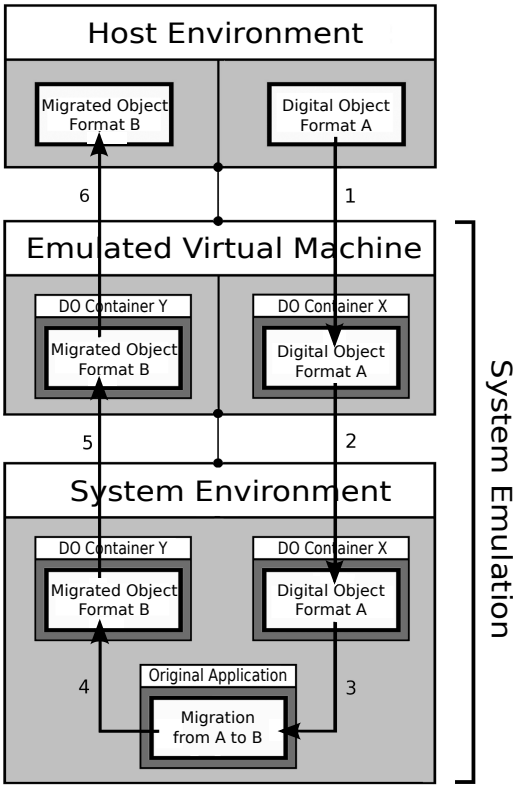


Figure 1: Migration-by-emulation workflow and involved system components.

- *Interactive Workflow Description*
  An abstract description of all interactive commands to be carried out in order to perform a certain migration. Such a description consists of an ordered list of interactive input actions (e.g. key strokes, mouse movements) and expected observable output from the environment (e.g. screen- or system-state) for synchronization purposes.

3.1 Base Layer – System Environments

A core requirement for the described workflows is the availability of the original system environment. This is the context in which the digital objects are rendered or run. It has to combine suitable hardware and software components so that the object's creation environment or a suitable equivalent can be reproduced. As many original system environments cannot be run on today's computer hardware, emulated hardware is used. Besides supporting the requirements of the original environment, the emulator has to be equipped with appropriate interfaces for automation and framework integration [20].

No matter which emulator is chosen, contextual information about the original environment in which the digital artifact was created is also required. For example, the answers to questions such as "Which operating systems is WordStar 7.0 compatible with?" are less obvious today than twenty years ago. To overcome this knowledge gap, the process of computing the actual needs for an authentic rendering environment is formalized by the concept of view path [19, 17].

3.1.1 Container Preparation

In order to perform the migration using the original tools in their original environments, it is necessary to provide these tools in order to operate on the digital objects of interest. The objects must be injected into the emulated environment from the actual one, migrated and, finally, derived objects produced by those applications must be extracted. Injection/extraction of data into/from the emulated environments is possible via emulated data storage devices, e.g. via floppy, hard disk, and CD-ROM drives or via virtual network connection if supported by the emulator and the operating system. These can be seen as gateways for binary data exchange between the real and emulated environments. Usually emulators support the emulation of at least one storage device.

For example, with QEMU it is possible to activate the emulation of a floppy drive with a virtual floppy disk using the following argument pair: `qemu -fda floppy.img`. Here the file floppy.img refers to a virtual floppy disk prepared in the actual working environment of the user. The file contains data in the form of files subject to injection into the emulated one. Any modifications performed on the virtual floppy disk inside the emulated environment will be reflected in the corresponding image file. It can afterward be mounted onto the filesystem of the actual system in order to acquire the modified data.

Floppy disks are standardized for a wide range of different computer platforms and are a comparatively simple solution, but they are limited in size. An alternative method of data injection/extraction is the use of hard disk drives. They are modifiable and their size can be adjusted as desired. This allows for the injection of large quantities of DOs into the emulated environment. The production of empty hard disks can be automated, similar to the process of creating floppy images. First, the desired amount of storage space for the disk image is allocated. In a second step, a disk partition with a suitable file system has to be created. The first partition in the hard disk starts at 32256 byte offset. Finally, starting from the aforementioned offset, the allocated storage is formatted with the required file system.

3.2 Migration Component

The *migration component* (MC) is the main module visible to the end user, by exposing a simple *migrate* interface for (possibly) complex DO migration from format fmt_A to format fmt_B. The user requests a migration by providing a (set of) digital object(s) to be migrated, the requested final format, and a set of parameters. These parameters may restrict the migration path length set quality or cost criteria for the migration process. Based on these criteria individual migration steps are identified. Figure 2 illustrates the general mode of operation of a MC.

In order to ensure the remote-accessibility requirement, this component is to be implemented in the form of a Web service. Its integration into existing DP frameworks would allow its usage both as a stand-alone tool and as a part of more sophisticated preservation workflows. The PLANETS Inter-

operability Framework (IF) [8] is a suitable candidate for the integration requirement. In this framework each preservation tool can be invoked according to one of the predefined code interfaces. The chosen interface depends on the tool's role in the scope of the DP (e.g. object migration, characterization, viewing, comparison).

On a low level, the migration component would then represent the implementation of the PLANETS IF *migrate* interface. The migration-by-emulation Web service could be invoked by using it's WSDL description file.

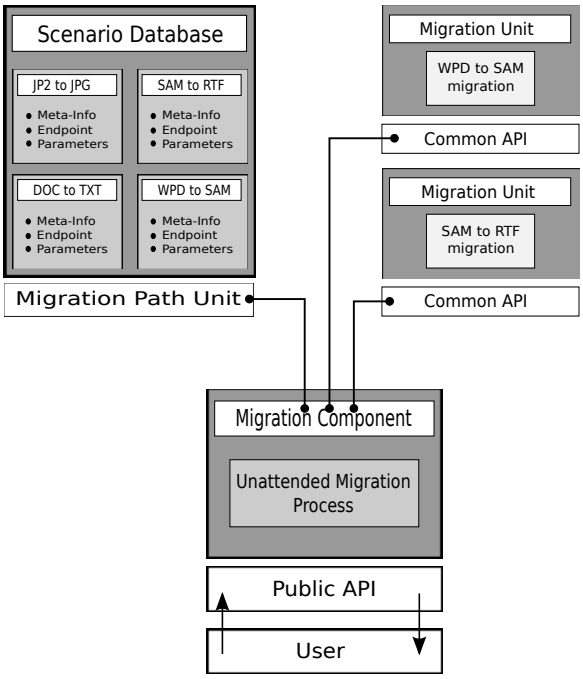


Figure 2: Migration component overview.

Based on the resulting identified migration path, the MC instantiates each node as a single migration unit. Beside this, the MC takes care of intermediate results and, if necessary, error reporting and recovery.

3.3 Migration Scenario

A migration scenario database describes atomic units for migration from format ftm_A to fmt_B and maintains available information about all atomic migration scenarios. "Atomic scenario" refers to a scenario not involving intermediate migrations. Such a scenario consists of a Web service endpoint and necessary meta-information (e.g. input/output formats, efficiency level, single/multiple DO support, author, timestamp). The input/output format identifiers need to be defined according to a conventional format registry system (e.g. Pronom[4]). Optional parameters may be required at instantiation time.

More complex format migrations cannot be carried out in a single step. Usually, several intermediate steps are required,

[4]The technical registry PRONOM, *http://www.nationalarchives.gov.uk/pronom*

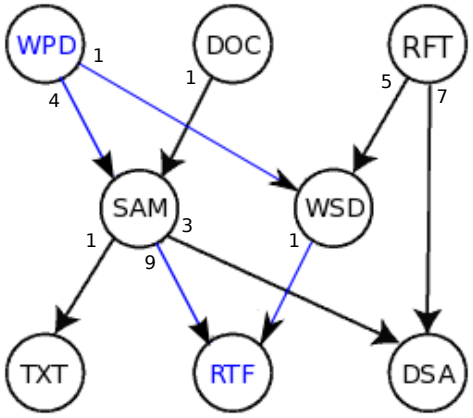


Figure 3: Migration graph. Nodes depict file formats, edges depict cost or quality weights.

but also several distinct paths are available between fmt_A and fmt_B. Hence, the scenario database can be represented as a directed graph with nodes representing supported file formats and edges describing weights, for instance based on the resulting quality or cost of a specific format migration. Figure 3 shows a possible migration graph between various text document file formats.

Depending on the migration requested by the user, the complete migration paths corresponding to it are to be calculated, being formed from the atomic ones. At this step the path preferences specified by the user in the parameter list can be taken into account. After the calculation, the component receives the path in the form of ordered lists of atomic scenario identifiers, which then will be instantiated and controlled by the migration component. In a distributed sce-

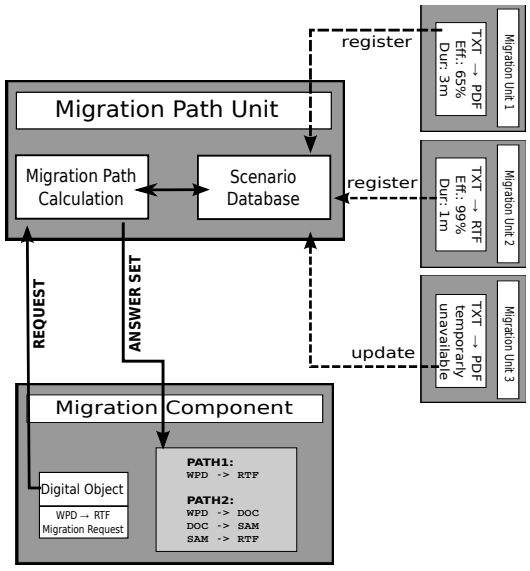


Figure 4: Migration path generation and calculation.

nario, different sites can register migration units in a centralized scenario database. Furthermore, based on recent runtime results and/or user feedback, quality measures like authenticity, reliability and runtime costs can be updated in the database (illustrated in Fig. 4).

For instance, a WPD (Word Perfect 6.0) to RTF (Rich Text Format) migration is requested. A user invokes the migration component by providing the required input files: an archive of DOs, input format – WPD, output format – RTF, and a parameter defining a maximum path length of two atomic migration steps. The migration component queries the scenario database through the migration path unit and receives a set of different migration paths. Among them only one path satisfies the maximum migration path length condition: WPD to SAM and SAM to RTF. First, a WPD to SAM migration is performed. The endpoint defined in the corresponding bundle is used to invoke the responsible migration unit. The meta-information of the bundle contains the field indicating that this service is able to perform the migration of multiple DOs in one operation.

Additionally, migration scenarios which produce two different output objects of different formats from one original artifact would be possible. Therefore, the DOs can be passed in an archive container (e.g. ZIP). After a migration an archive of migrated DOs is returned along with the success status.

According to the meta-information describing the second migration unit (SAM-to-RTF), the service is unable to perform the migration of multiple DOs in one turn. Therefore, the archive of intermediate DOs is extracted by the migration component to a local temporary storage. From there the intermediate DOs are migrated one by one and the results of the migration are retrieved. The resulting DOs in RTF format are packed into an archive and returned to the user.

If a representative test set of input and output files are kept for each migration scenario, software updates or other minor changes on the emulated system environment can be tested in an automated way for compatibility.

Such migration scenarios can then be registered with the preservation framework as simple migration services.

3.4 Migration Units

The combination of emulator and original environment is itself not sufficient for automated migration workflows. It just provides the base to view, modify or migrate a digital object manually. Thus, a *migration unit* (MU) combines the system environment with an interactive workflow description (IWD). The IWD is an abstract description of a recording of all user interactions of a specific task. Such interactions of a human user with the computer UI of the original applications is represented by a series of input events, such as mouse clicks/movements and key strokes. These events can be simulated using remote desktop control systems like VNC using an application like VNCplay [25], allowing them to be performed in an unattended manner [10].

3.4.1 Workflow Recording

To create a specific MU the original system environment is to be combined with an IWD describing the required user interaction. This is realized by running actions corresponding to manually prerecorded migrations, for instance a WPD-to-RTF conversion using the original application WordPerfect 6.0 in Windows 3.11. The input events produce the following actions in the emulated environment:

1. The WordPerfect 6.0 is executed via a mouse click on its icon.
2. The "Open" menu of Word Perfect 6.0 is chosen.
3. In the opened dialog box, the DO on the attached hard disk drive is chosen and loaded.
4. The "Save As" menu of Word Perfect 6.0 is chosen.
5. In the opened dialog box the new file name is chosen according to the conventional naming scheme and the DO is saved in RTF format on the same hard disk. The migrated object is produced at this point.

Fig. 5 illustrates the workflow of a migration unit creation. These or similar workflows are recorded by the *scenario preparation component* to generate abstract interactive workflow description for specific user interaction. A more complex example could be setup as follows: An original environment contains an AMI Pro 3.1 installed on Windows 95 with a virtual Postscript printer running in QEMU. This allows the recording of several different scenarios:

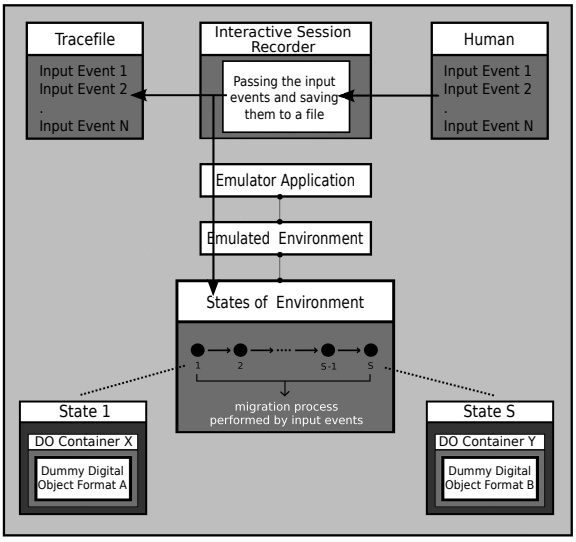


Figure 5: Workflow of creating a complex migration unit.

- Lotus AMI Pro to TXT
- Lotus AMI Pro to RTF
- Lotus AMI Pro to Postscript (PS) using a virtual printer
- Combinations of the aforementioned atomic migration steps, to produce two different objects from one input, like TXT and PS

- Non-native formats by different input filters available from AMI Pro

The last one might help to deal with unavailable applications for certain deprecated formats or could provide a base for rendering comparisons of other tools. All the mentioned options use exactly the same base environment.

The workflow recording is then kept as metadata for a certain stored system environment which is provided by some backend store keeping the software archive [21, 16] of original environments. There could be several recordings registered with one original environment. In theory – as this has not been thoroughly tested yet – the emulator should be exchangeable with a computer architecture compatible type, as the original system should perform roughly the same way on the other emulator. This was tested for a chain of updates over a number of different versions of QEMU [22].

3.4.2 Workflow Replication

The recorded set of actions is later executed by the MU for each DO in an automated manner. After sending the last input event and observing its expected outcome, the interactive workflow replication service finishes its work. The workflow replication service sends a request for the termination of the emulated environment, using the current session identifier. The emulation service frees all resources, detaches the hard disk image with the newly migrated DOs and returns it to the workflow replication service.

The operating system and the original tool are used: in this case Windows 3.11 and WordPerfect 6.0 are installed. According to the workflow, the migration is invoked by providing it with the archived DOs and optional parameters. The workflow replication service parses the optional parameter list and acquires the references to the three necessary objects: a suitable system environment, the interaction workflow description and the target emulator. It then extracts the received DOs from the archive to the local storage, while calculating their total size. The DOs are renamed according to the predefined naming scheme and the correspondence between the real and virtual names is stored separately. The next step is the DO container creation and injection of data.

The DO container creation method is invoked and information regarding the desired filesystem of the hard disk drive and its size is provided. The size is set to a value large enough to hold the DOs and the resulting output files, which will be produced in the emulated environment. The DO container creation subunit acquires the input data and produces the hard disk image file. It then injects the DOs without changing their filenames. After the operation is completed, it returns the prepared container to the workflow replication service.

When the DO container is prepared, the emulated system environment can be started. The workflow replication service invokes one of the emulation service endpoints stored in the scenario database and one of the suitable emulator associated with it. It then invokes the respective emulation service by providing the following: the emulator ID, the operating system image of the system environment object, any secondary objects and/or directives necessary for the emulation of this operating system, and the prepared container with the DOs in it.

The system emulation service receives the input data and starts the appropriate emulated system environment. It also attaches the hard disk container with the DOs. The remote desktop control is activated. The system emulation service returns a success message with a port number for connection as well as the session identifier. The workflow replication service then initiates the replication of an interactive session.

4. DP-FRAMEWORK INTEGRATION

The Java-based migration-by-emulation service prototype provides all main components. Those include basic migration units, migration component services as well as a simple scenario database. A further challenge is the integration of the developed components in standard DP-frameworks (e.g. PLANETS) and their associated workflows. The specific objectives to be pursued involve enabling interactive user access to original system environments to create new environments, create or modify existing ones but also to create abstract interactive workflow descriptions. Having original environments and IWD for certain file types, these migrations have to be made available as simple steps of more complex DP-workflows.

4.1 Preparation User Interface

The PLANETS *view* Web service interface is designed to render an arbitrary digital object. The service takes a digital object and returns an URI which points to the rendered result. If the digital object requires a running rendering engine, the service offers methods for querying the engine's state and allows sending commands to it. This interface is suitable to integrate the user preparation and recording workflows within the PLANETS framework.

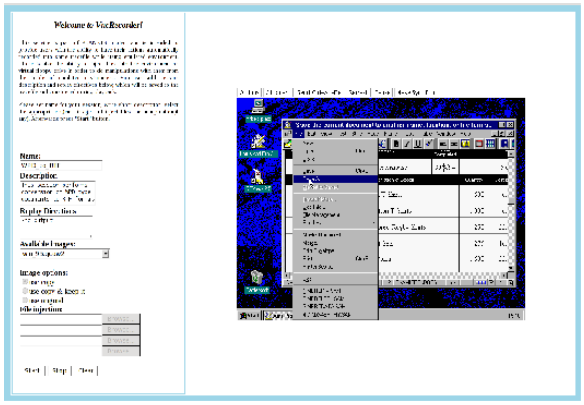


Figure 6: Grate-R interface to the PLANETS service for interactive workflow recording.

Preconfigured original environments are deployed and the Grate-R (cf. Section 2) VNC record service is used to generate an abstract interactive workflow description. They were then attached by hand to the appropriate original environment to form a proper migration unit.

4.2 Testbed Migration Workflows

The user interface to existing migration-by-emulation workflows is realized via the standard PLANETS testbed GUI [14, 1]. The services are called from within the testbed standard procedures. As migration-by-emulation services should be accessible the same way as standard command line tools they are registered and deployed using the same methods within the testbed. Thus, the testbed user interacts with the Web interface in the usual way.

Since there is no suitable PLANETS wide tool registry available supporting complex view path calculation, the migration path computation is not part of the framework yet. The PLANETS testbed was originally designed to retrieve the available (migration-)services from the PLANETS service registry. Each deployed service registers itself and describes its capabilities (here migration paths) using a `describe()` method. However, migration-by-emulation services can become more complex then input-output oriented migration tools or methods (e.g. complex migration graph). The *migration component* was designed to hide some complexity from the user e.g., automated complex path selection based on cost parameters provided by the user.

Due to the construction of the PLANETS framework, either single step migrations or preselected complex migration paths can be exposed to the testbed users. More complex migrations have to be constructed within the framework's capabilities. To provide all migration options to the PLANETS testbed user, a dynamic `describe()` method can be used, retrieving all available endpoints from the scenario database. The coupling of atomic migration units depends on the user for now.

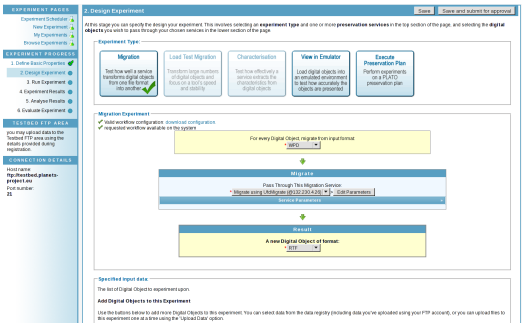


Figure 7: Selecting the Word Perfect to RTF migration in the Testbed.

4.3 Experiments and Results

Two different migration services were registered within the testbed. One of them is a truly atomic migration accepting WPD as input and producing RTF as output. Thus the resulting file is directly delivered to the user after the procedure succeeds (Fig. 7). The second service is more complex as it takes an AMI Pro text document (SAM) as input and produces two different outputs, a TXT and a PDF. The TXT is the result of a classical "save-as" migration. The PDF is generated by sending the document to a virtual printer generating PS as output. This file is then loaded to the Ghostview application, which renders a pdf from it.

Additionally, a virtual disk handling service was programmed to produce disk image containers for different emulators with the option of specifying a range of supported filesystems understood by the original system environments. The creation of a QEMU compatible container with a FAT filesystem is comparably simple, other containers and filesystems are supported as well by using the "qemu-img" container conversion tool.

The WPD to RTF service deploys the original DOS Word Perfect 6.0 application running in QEMU using a Windows 3.11 environment with mouse and keyboard interaction (Fig. 8). The procedure was tested with a small number of different WPD files.

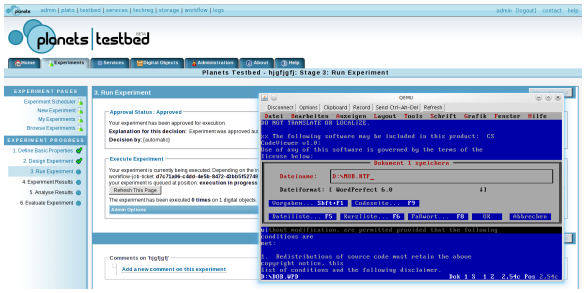


Figure 8: Running Word Perfect migration in Testbed. QEMU screen output attached for monitoring purposes.

The second migration unit was deliberately made to be more complex. We wanted to demonstrate the feasibility of producing more then one output file from a single input. This helps to evaluate and compare different workflows regarding runtime, reliability and complexity. The experiment was run on a set of a couple AMI Pro files containing between three and 15 pages. The average runtime took 2 minutes 33 seconds to complete. Of course it would have been possible to execute the PS to PDF migration by a simple command line tool in another atomic migration. But we wanted to stress the interaction playback component on a longer running workflow involving more than a single application. The procedure did not fail in all experiments we ran on the testbed.

After invoking the migration service locally the same test was repeated by calling the service over the Internet from some geographically different PLANETS instance. It ran successfully and the different execution time difference for the conversion was insignificant.

Due to limited capacity in time and the small number of available objects for testing, the evaluation of the migration services was yet short. Thus the figures collected need still to be proven for a larger number of objects running in migration processes on different machines.

A major drawback in all experiments was the availability of a proper set of test files. Such a ground truth set of files enriched with descriptive metadata on features like length, containing pictures, special font sets, complex layouts and other features would be very helpful to evaluate and compare preservation workflows and migration tools.

4.4 Next Steps

Still a problem is the black box character of the running migration unit (cf. [11]). In the present state of development, not much feedback on the state of active migration workflows is available and reported back to the user. While it is generally possible to attach another VNC viewer to the running emulator the migrate interface lacks any methods for querying its execution status in run-time. Thus, research at Freiburg University focuses on the VNCplay tool and the files used by it. The IWD file could be exploited to gather information on the actual state of an ongoing migration. Every interaction and expected screen result relates to a certain state of the process. This could be used to generate progress information or produce more meaningful error messages. Having this in place, the runtime of the entire process should be optimized by eliminating redundancy in the IWDs.

In future steps the authors hope to enhance the service to handle more emulators beside QEMU and include additional non-x86 original environments. Plus, more runtime improvements like system resume-restart or multi-object migrations will be looked into. The additional experience and information should help for a better preservation planning by providing quality metrics and cost estimations [3].

There are interesting alternative approaches like Polyglot for automated file conversions. The service was developed by [15] originally to convert 3D model files into different formats. The underlying concept is quite different to the presented one and is worth comparing and benchmarking.

5. CONCLUSION AND OUTLOOK

Migration-by-Emulation services allow a wide range of different file format conversions. Compared to simple command line migration tools those services are more complex to setup and deploy. Nevertheless after having the workflow established a new migration scenario is simpler to be integrated. Often only a new workflow is to be recorded to add another service like the conversion of WPD to TXT or Postscript. Even if a completely new input like Wordstar files are to be supported, only an appropriate original environment is to be extended without the requirement to program a new tool wrapper. Thus, the system presented allows to easily compare the migrations run on different original environments.

The operation and management of migration-by-emulation services as presented could be decentralized and several institutions could share the workload or specialize on certain environments and share their expertise with others.

Our implementation focused on the feasibility of the preservation framework integration. Future research is dedicated to the speedup of workflows by looking into the VNC recording and playback. The interactive workflow descriptions are a good starting point for optimization. They could be enriched with additional metadata to use it for progress reporting. A certain state in the metadata directly correspondents to a state of the migration workflow and could be reported back to the preservation framework.

The interative workflow description could be modularized to better identify the different stages, like original operating system booting, application starting, artifact loading, and saving in a new format. This information could not only be used for feedback but to identify checkpoints. Those checkpoints could help with error recovery for restarting the procedure after failed attempts. Additionally, these workflows could help to evaluate future versions of emulators before they get integrated into preservation systems [22]. These issues are part of the ongoing research at Freiburg University.

Depending on the future needs arising from the requirements of Migration-by-Emulation, a couple of different strands seem to be worth exploring in the coming years. Especially the splitting of the preservation application into a simple user's front-end and an easy-to-extend server backend is very attractive, as it could be adapted to many preservation framework services. Nevertheless, future emulator research and development should take the "preservation-awareness" regarding automation and long-term support more into consideration. With the ongoing research in the KEEP[5] or SCAPE[6] projects new solutions and progress in wrapping emulators and handling large scale migrations could be expected. Beside this, new insight into emulation metadata and tool registries [2] will help to automate more steps of the migration-by-emulation workflow. Nevertheless, a number of challenges like software archiving, emulation knowledge base and a proper definition of a test set of digital artifacts remain open.

6. REFERENCES

[1] B. Aitken, S. Ross, A. Lindley, E. Michaeler, A. Jackson, and M. van den Dobbelsteen. The planets testbed – a collaborative research environment for digital preservation. In M. Lalmas, J. M. Jose, A. Rauber, F. Sebastiani, and I. Frommholz, editors, *Research and Advanced Technology for Digital Libraries, 14th European Conference, ECDL 2010, Glasgow, UK, September 6-10, 2010. Proceedings*, volume 6273 of *Lecture Notes in Computer Science*, pages 401–404. Springer, 2010.

[2] 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.

[3] C. Becker, H. Kulovits, M. Kraxner, R. Gottardi, A. Rauber, and R. Welte. Adding quality-awareness to evaluate migration web-services and remote emulation for digital preservation. In *Proceedings of the 13th European Conference on Digital Libraries (ECDL09)*, 2009.

[4] V. Delwardia, S. Marshall, and I. Welch. Experiments in Remote Mobile Gaming. In *AUIC: Australasian User Interface Conference*, 2009.

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

[6] E. Genev. VNC Interface for Java X86-Emulator Dioscuri. Online, http://hdl.handle.net/10760/15102,

[5]See the KEEP emulation framework information plus development pages at *http://www.keep-project.eu* and *http://emuframework.sourceforge.net*

[6]See project pages at *http://www.scape-project.eu*

October 2010.

[7] 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.

[8] R. King, R. Schmidt, A. N. Jackson, C. Wilson, and F. Steeg. The planets interoperability framework. In *Proceedings of the 13th European Conference on Digital Libraries (ECDL09)*, pages 425–428, 2009.

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

[10] K. Rechert and D. von Suchodoletz. Tackling the problem of complex interaction processes in emulation and migration strategies. *ERCIM News*, (80):22–23, 2010.

[11] K. Rechert, D. von Suchodoletz, and R. Welte. Emulation based services in digital preservation. In *JCDL '10: Proceedings of the 10th annual joint conference on Digital libraries*, pages 365–368, New York, NY, USA, 2010. ACM.

[12] K. Rechert, D. von Suchodoletz, R. Welte, M. van den Dobbelsteen, B. Roberts, J. van der Hoeven, and J. Schroder. Novel workflows for abstract handling of complex interaction processes in digital preservation. In *Proceedings of the Sixth International Conference on Preservation of Digital Objects (iPRES09)*, 2009.

[13] T. Richardson. The rfb protocol. Online, *http://www.realvnc.com/docs/rfbproto.pdf*, 2009.

[14] R. Schmidt, R. King, A. Jackson, C. Wilson, F. Steeg, and P. Melms. A Framework for Distributed Preservation Workflows. *International Journal of Digital Curation*, 5(1), 2010.

[15] University of Illinois at Urbana-Champaign, NCSA. Towards a universal file format converter. Online, *http://isda.ncsa.uiuc.edu/NARA/conversion.html*, 2011. Online resource.

[16] M. van den Dobbelsteen, D. von Suchodoletz, and K. Rechert. Software archives as a vital base for digital preservation strategies. Online, *http://hdl.handle.net/10760/14732*, July 2010.

[17] J. van der Hoeven and D. von Suchodoletz. Emulation: From digital artefact to remotely rendered environments. *International Journal of Digital Curation*, 4(3), 2009.

[18] 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.

[19] D. von Suchodoletz. *Funktionale Langzeitarchivierung digitaler Objekte – Erfolgsbedingungen für den Einsatz von Emulationsstrategien.* Cuvillier Verlag Göttingen, 2009.

[20] D. von Suchodoletz. A Future Emulation and Automation Research Agenda. In J.-P. Chanod, M. Dobreva, A. Rauber, and S. Ross, editors, *Automation in Digital Preservation*, number 10291 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2010. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Germany.

[21] D. von Suchodoletz. Das Softwarearchiv – Eine Erfolgsbedingung für die Langzeitarchivierung digitaler Objekte. *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare*, (63):38–55, 2010.

[22] D. von Suchodoletz, K. Rechert, and A. N. Tchayep. QEMU – A Crucial Building Block in Digital Preservation Strategies. In W. Müller and F. Pétrot, editors, *1st International QEMU Users' Forum – DATE 2011 Workshop*, Grenoble, France, 2011.

[23] D. von Suchodoletz, K. Rechert, R. Welte, M. van den Dobbelsteen, B. Roberts, J. van der Hoeven, and J. Schroder. Automation of flexible migration workflows. *International Journal of Digital Curation*, 2(2), 2010.

[24] R. Welte. *Funktionale Langzeitarchivierung digitaler Objekte – Entwicklung eines Demonstrators zur Internet-Nutzung emulierter Ablaufumgebungen.* Südwestdeutscher Verlag für Hochschulschriften, 2009.

[25] N. Zeldovich and R. Chandra. Interactive performance measurement with vncplay. In *ATEC '05: Proceedings of the annual conference on USENIX Annual Technical Conference*, pages 54–64, Berkeley, CA, USA, 2005. USENIX Association.

Emulation as a Business Solution: the Emulation Framework

Bram Lohman
Tessella
President Kennedylaan 19
Den Haag, The Netherlands
bram.lohman@tessella.com

Bart Kiers
National Library of the Netherlands
Postbus 90407
Den Haag, The Netherlands
bart.kiers@kb.nl

David Michel
Tessella
26 The Quadrant
Abingdon, United Kingdom
david.michel@tessella.com

Jeffrey van der Hoeven
National Library of the Netherlands
Postbus 90407
Den Haag, The Netherlands
jeffrey.vanderhoeven@kb.nl

ABSTRACT

Emulation is often considered a technically very complex subject. The association with complexity has long prevented it from being considered in an end-to-end business solution for long-term preservation and access to digital collections.

The Emulation Framework solves this problem by automating the steps required to render an unknown digital object in its original environment: characterising the object to determine its type; determining the environment required to render that type of object; setting up the required software and emulators providing the hardware; and configuring the environment to properly render the digital object. Automating these steps allows a novice user to easily render a digital object in an environment for accessing it in its original form.

Each of the four steps of the emulation workflow are described in detail, providing a simple tool for managing a complex decision making process.

Keywords

Emulation, framework, digital preservation, workflow, business solution, KEEP, characterisation, technical environment, viewpath, software

1. INTRODUCTION

Long-term preservation of digital objects not only implies looking after their conservation, but also necessitates the development and execution of strategies to ensure these objects remain accessible and understandable in the future.

iPRES2011, Nov. 1–4, 2011, Singapore.

The KEEP [6] (Keeping Emulation Environments Portable) project is a research project co-funded by the European Union under the Seventh Framework Programme (FP7). It does research into media transfer, emulation and portability of software from a technical and legal perspective [15]. The project extends previous work on emulation, such as the Dioscuri project that developed an x86 emulator [1], and the Planets project which amongst others created emulation and migration services [8]. Emulation is a vital strategy for permanent access, but it requires several more steps to become mature [2]. KEEP aims to deliver a strategy that provides permanent access to multimedia content (such as computer applications and console games), not only now but also in the long term.

2. WHY EMULATION?

Emulation recreates a computer environment (target) on top of another computer environment (host) [11]. It is a proven technology that can be used to cope with obsolescence of hardware and software. By rendering a digital object within an environment running original software, an authentic recreation of that object in its native computer environment is given. The advantage of such a strategy is that no change to the digital object is required which offers better conditions for displaying it in its original form. Another advantage of emulation is that it also works for complex digital objects such as software applications (e.g. games), websites or visualisations of data sets.

3. AN END-TO-END EMULATION WORKFLOW

One of the main problems facing emulation is the lack of knowledge in identifying and configuring the technical environment required to render a digital object. The KEEP project recognises this issue, and in May 2011 released the Emulation Framework (EF), an open source solution for applying emulation as an access strategy for files and computer programs. It is released under the open source Apache 2.0 license and is freely available [3]. When a user requests an item from a digital collection and this item requires an ob-

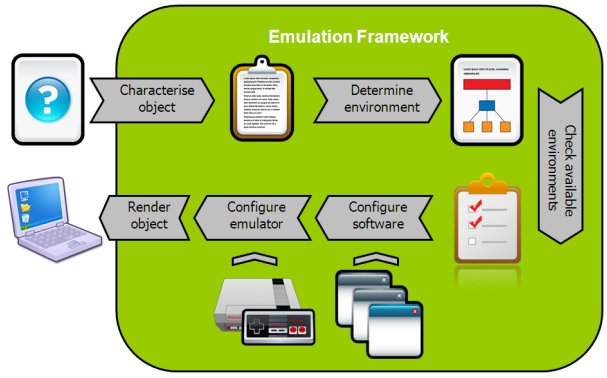


Figure 1: Emulation Framework workflow.

solete computer environment to render, the EF offers a solution that does not require any in-depth knowledge, following the workflow steps depicted in Figure 1.

The EF automates the identification of an (unknown) digital object; the need to know what application, operating system (OS) and hardware is required to emulate the object; preparing the selected environment for use; and configuration of the environment for rendering the digital object. These four steps are explained in more detail in the following sections.

3.1 Characterising an unknown digital object

Characterisation is a subject in digital preservation that has been researched in depth. This research has resulted in several tools that can characterise an unknown digital object, i.e. determine its file format. Harvard University Library Office for Information Systems released a tool, called the File Information Tool Set (FITS) [4], which acts as a wrapper for several proven open source tools. FITS identifies, validates, and extracts technical metadata for various file formats. It normalises, consolidates, and reports any errors in the output of the wrapped tools. FITS currently uses Jhove, National Library of New Zealand Metadata Extractor, DROID, FFIdent, Exiftool and File Utility [4]. It was an obvious choice to use this tool for characterisation in the EF.

The tools have no problem identifying the top 10 most common file types used in memory institutions [13]. Unfortunately, they lack support for most objects used in the emulation community: computer games, cartridges and disk image files created by that community. These disk images include, for example, common Amiga and Commodore 64 formats. During EF development, support for some of these formats was added by reconfiguring the FITS tool.

The FITS software also provides a novel selection method: it returns the number of tools that agree on the determined file format. This can be used as a measure of success, along with validation, and is used within the EF to automatically select the digital object's file format. Once the file format has been identified, the next step is to select an environment that provides the dependencies to render it in its original context.

3.2 Determining a rendering environment for a known digital object

The EF defines a rendering environment in a similar way as a viewpath [14] (or emulation pathway), of which two examples are shown in Figure 2. This is a structured description of the complete hardware and software stack needed to render a digital object. For today's PC architectures it consists of four layers (digital object, rendering application, OS, hardware platform), although for other architectures, not all layers are required. Console games, for example, usually only have two: digital object and hardware platform (including embedded software).

This is the simplest approximation of a rendering environment. Although it works for simple cases, in practice the connections between layers are more complicated: file formats require certain versions of applications to render properly; integration of application and OS requires specifics such as drivers and libraries; integrating OS and hardware platform depend on specific firmware to work together. In the current design, these more complicated cases are not supported and the EF uses the simple four-layer model of digital object file format, application, OS and platform.

Keeping the model simple does have an advantage. As complexity is increased exponentially at every layer – for each format there is often more than one application to support it; each application can run on different OS's, and each OS usually supports many hardware configurations (or emulators) – a model with fewer layers has lower complexity.

Technical metadata links each of the layers, and thus a directed graph can be generated with the digital object file format as the root node. The EF relies on an internal database containing metadata but can also interface with technical registries such as PRONOM [9] or PCR [16] to retrieve this metadata. To address the lack of publicly available registries, KEEP is addressing this issue [12] as well.

The EF is currently prototyping a novel method of combining information from different registries to ensure more robust information is used to create technical environments.

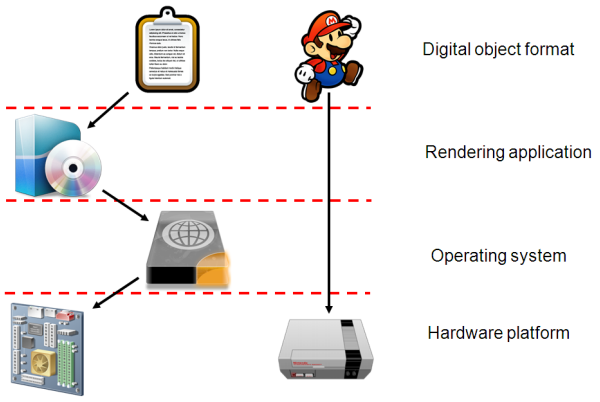


Figure 2: Environmental dependencies of digital objects.

3.3 Preparing the hardware and software stack

The main technical problem is merging the four distinct components defined by the digital object and its dependencies into one assimilated environment: to view a digital object, the bit stream has to be interpreted by an application, which in turn has to be configured specifically (i.e. installed) on an OS that is configured for a particular piece of hardware. At rendering time, the stack is difficult to view as individual components.

There are several approaches to generate an environment:

- Use an automated method to merge the four components at runtime.
- Prepare a complete environment beforehand to be used.
- Use a combination of these.

Although for simpler environments, such as MS-DOS, the 'merging' step can relatively easily be automated, software and hardware systems have in recent years become increasingly complex. Environments running on today's desktop are based on customised hardware running a specifically set up OS that has applications that are tightly coupled to it (e.g. registry entries, library versions, etc.). Setting up such an environment requires a high number of complex choices to be made. The problem is not so much that it can not be done, but there are so many corner cases and exceptions, that the effort of creating an automated method that can reliable generate any selection of environments far exceeds the benefits. The University of Freiburg is continuing research into this area [17].

The second approach requires setting up the OS, application and digital object as required by the selected environment in advance. The only reliable way is a human initiated, time-consuming process, but it only needs to be done once for an environment; it can then be stored and accessed whenever required.

The EF has tested the third approach as a proof of concept. In general, the digital object (top layer) and the hardware platform (bottom layer), are only loosely coupled to the layers in between. Those layers, the application and OS, however, are so interdependent, that only by setting these up beforehand can be guaranteed that it is done correctly. To address this, the EF created a 'Software Archive', a web service that holds prepared application/OS disk images along with metadata containing details of the OS, applications and hardware requirements. Using technical registry metadata, an appropriate disk image can be selected from the database that fulfils the environmental requirements.

Similarly, a separate web service, the 'Emulator Archive', holds the emulators that can be used to represent the hardware. It also contains metadata to match the required hardware selected in the technical environment, along with the type of software image it supports, and thus a match can be made between the emulated hardware and the OS/application layer.

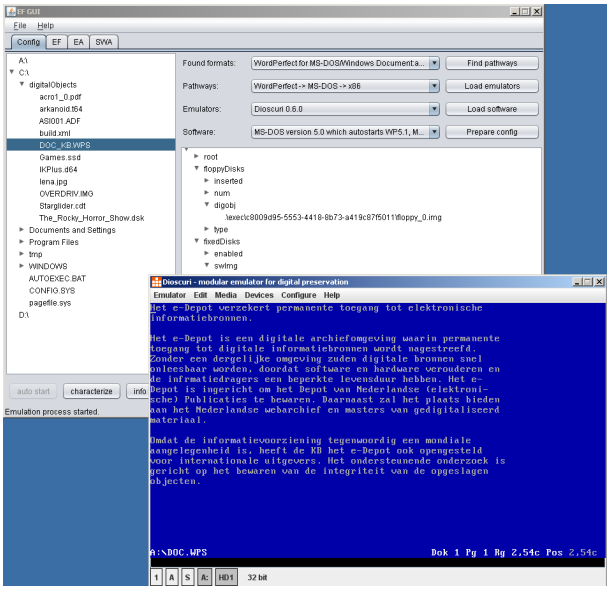


Figure 3: The Emulation Framework rendering a digital object in its original environment.

3.4 Configuring the environment to render the digital object

Configuration of a hardware platform, despite there being many different variations, can be made much simpler by creating a high-level hardware component model of it. Making the model sufficiently generic allows it to be used for multiple emulators. Although the low-level details for each hardware set may be different for each emulator (even if they address the same platform!), using a single model greatly simplifies the problem. The EF currently manages to configure 7 emulators covering 6 platforms using a single abstracted model.

To generate an emulator-specific configuration, the EF makes use of a template processor, a software component designed to combine a data model with a template to produce a result document [5]. Each emulator specific template contains the grammar for configuring that emulator, which when combined with the emulator-agnostic data model, generates the emulator specific hardware configuration that the software requires. Given the configuration options for the environment (such as number and type of floppy drives, hard disk parameters, CPU settings, etc.) a customised configuration can be created for each emulator.

The last part of this step is providing the digital object to the application within the disk image. Because the application and OS disk image is prepared prior to the process, the digital object cannot easily be inserted into it. However, it can be attached to the emulated hardware as a separate disk image that with the right configuration can allow the application within the disk image to access the digital object. For example, a disk image containing MS-DOS and WordPerfect is provided to an x86 emulator as a hard-disk, and a floppy disk image containing the accompanying WordPerfect file is also provided to the platform. Within the rendering environment, it is possible to boot the OS from the hard-disk, start the application, and read the digital object from the

attached floppy disk from within the application.

This completes the last step of the workflow to render an unknown digital object in its original environment, as can be seen in Figure 3.

4. BUSINESS INCENTIVES

With the release of the EF using emulation tools has become more accessible, bypassing difficult setups or technical restrictions. The EF runs on Java, making it compatible with all mainstream computer platforms. Moreover, management of required emulators and software packages has become more organised by using the service-oriented Emulator and Software Archives. With the large number of freely available emulators, most computer platforms can be emulated by including them in the EF. However, care must be taken when using old applications and emulators as software licenses and hardware patents can restrict limitations of use [15]. For this reason the current release of the EF only uses open source emulators and applications.

5. ONGOING RESEARCH

Building on the first release of the EF (May 2011), the KEEP project is working on improving the software. Two new releases are planned before the end of the KEEP project in February 2012. These will incorporate the user feedback from tests performed by the Bibliothèque nationale de France, Dutch National Archives, Computerspiele Museum Berlin, research institute CERN and the Netherlands Media and Art Institute. Altogether these organisations represent five major domains: library & archiving, culture, research and art.

Furthermore, KEEP is doing research into remote emulation, with the goal of accessing the rendered digital environment from a thin client. This will move the high requirements emulators place on the underlying hardware to a server specifically built for the task.

6. CONCLUSION AND BENEFITS

The EF has shown that an end-to-end business solution using emulation to render a digital object in its original environment is feasible. The EF is currently freely available, allowing individuals and institutions to take advantage of the possibilities to unlock their digital archive to the wider public at a very low total cost of ownership. Especially those digital objects for which migration, currently the main digital preservation strategy, provides no accessibility is this solution an alternative.

Ongoing pilots at the National Library of the Netherlands, the German Computerspielemuseum and integration with Tessella's Safety Deposit Box [10] show that in a wide range of environments the EF offers access to a large set of digital objects. With the Open Planets Foundation [7] in place, a platform exists that will ensure this solution continues to be developed, and also guarantees continual support.

All in all, the EF offers an effective way of ensuring long-term access to practically any digital object, and can be put to use by any user or institution regardless of the technical knowledge available.

7. ACKNOWLEDGMENTS

KEEP has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no ICT-231954.

8. REFERENCES

[1] Dioscuri — the modular emulator. `http://dioscuri.sourceforge.net/`. Accessed: 01-Sep-2011.

[2] Emulation Expert Meeting Statement. `http://www.kb.nl/hrd/dd/dd_projecten/projecten_emulatie-eemstatement-en.html`. Accessed: 01-Sep-2011.

[3] Emulation Framework. `http://emuframework.sourceforge.net`. Accessed: 01-Sep-2011.

[4] File Information Tool Set (FITS). `http://code.google.com/p/fits`. Accessed: 01-Sep-2011.

[5] FreeMarker — Java Template Engine Library. `http://freemarker.sourceforge.net`. Accessed: 01-Sep-2011.

[6] KEEP project. `http://www.keep-project.eu/`. Accessed: 01-Sep-2011.

[7] Open Planets Foundation. `http://www.openplanetsfoundation.org`. Accessed: 01-Sep-2011.

[8] Planets project. `http://www.planets-project.eu/`. Accessed: 01-Sep-2011.

[9] PRONOM — the online registry of technical information. `http://www.nationalarchives.gov.uk/PRONOM`. Accessed: 01-Sep-2011.

[10] Safety Deposit Box. `http://www.digital-preservation.com/solution/safety-deposit-box`. Accessed: 01-Sep-2011.

[11] What is Emulation? `http://www.kb.nl/hrd/dd/dd_projecten/projecten_emulatiewatis-en.html`. Accessed: 01-Sep-2011.

[12] D. Anderson, J. Delve, D. Pinchbeck, L. Konstantelos, A. Lange, and W. Bergmeyer. D3.3 final document analyzing and summarizing metadata standards and issues across Europe. Technical report, KEEP project, September 2010.

[13] S. v. Bussel and F. Houtman. Gap analysis: a survey of PA tool provision. Technical report, Planets project.

[14] R. Diessen and J. Steenbergen. Long Term Preservation Study of the DNEP Project. Technical report, IBM, National Library of the Netherlands, December 2002.

[15] J. v. d. Hoeven, S. Sepetjan, and M. Dindorf. Legal Aspects Of Emulation. *iPRES 2010 proceedings*, July 2010.

[16] L. Montague and S. v. Bussel. PLANETS Core Registry: Future Vision Document. Technical report, The National Archives, National Library of the Netherlands, May 2010. PLANETS project, PC3-D24.

[17] D. v. Suchodoletz, K. Rechert, J. Schroder, and J. v. d. Hoeven. Seven steps for reliable emulation strategies solved problems and open issues. *iPRES 2010 proceedings*, July 2010.

Design Decisions in Emulator Construction: A Case Study on Home Computer Software Preservation

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

Preserving software is widely recognized as a far more complex task then preserving static data. Emulation is usually the chosen preservation action to enable the execution of programs of obsolete systems. In this work we show how software extracted from obsolete media was preserved by developing an emulator. We explain the reengineering work involved and the design decisions made as well as the options for data injection into and extraction from the emulated environment.

In previous work, data and programs stored on audio tapes were extracted and the resulting audio files were transformed into digital objects. The objects retrieved were mainly programs, requiring emulation for execution. As no emulator for the original system previously exists, we here show how we implemented one. We first describe the system in more detail and explain the reengineering of the view-path for the execution of programs on the original system. We show how an existing emulator for a video game system was expanded by emulation capabilities for the view-path of the home computer and how the different options for data exchange with the host environment were implemented on different levels in the view-path. We explain how differences in input and output formats and methods influence the development of an emulator and that, depending on the original system, the transfer of data between the emulated environment and the host environment enforces implicit migration of the data to become usable.

1. INTRODUCTION

Preserving digital objects for a long term does not only concern preserving static data like pictures or text documents. For a wide range of digital objects not only data has to be preserved but the actual rendering process of data is significant. This is especially true, when a digital object has to be continuously rendered, as in the preservation of software. But also whole business or scientific processes need to be stored for a long term to be able to exhume them at a later time and run them in a changed environment. One of our main concerns for preserving processes is keeping them accessible and the software originally used executable.

Preserving software across rendering environments, i.e. executing the software on a platform it was not designed for, is usually solved by executing the software in an emulator emulating the hardware of the platform and running on a different host platform. While the advantage of a hardware emulator is that it can potentially run all software designed for the hardware it emulates, it is a quite complex task to build an emulator [5] and involves expert knowledge about the hardware specifications of the original system. It is also necessary to not only emulate the hardware, but also to provide methods for providing input to the emulated system, either in the way of interaction with the system by using keyboards or other input devices, but also by injecting data from files into the system. Extracting data for usage in the host environment is also an important issue not tackled by most emulators today.

As previously published in [7] we extracted data encoded in audio wave forms from cassette tapes. Almost all the data extracted was programs written in a dialect of the computer language BASIC. The programs where converted from their original binary form to source code in readable text format. As preserving the source code is only the first step of preserving the programs, research on potential rendering environments was carried out. In this paper we now demonstrate the development of an emulator for the system and show which design decisions have to be made and what problems one has to deal with even with a fairly simple computer architecture. We show what one must consider so an emulator developed can be used for digital preservation by providing functionality for injecting data into the emulated environment and extracting data for use on the host system.

This paper is structured as follows. First we provide related work relevant for this paper. In Section 3 we examine the view-path of the original system and provide information on how the different components involved interact. We present how we implemented the view-path in an emulator in Section 4. In Sections 5 and 6 we explain the reengineering work necessary for data exchange between the emulated environment and the host environment. We explain what choices we were given to solve certain problems and what design decisions were taken for implementing the functionality, keeping digital preservation in mind. Next, we show

iPRES2011, Nov. 1–4, 2011, Singapore.

how the image rendered by the emulator can be evaluated against the original system and other alternative rendering environments. In Section 8 we then discuss other possible preservation actions besides emulation on different levels in the view-path. Finally we show our conclusions and give an outlook to future work.

2. RELATED WORK

Preserving software for obsolete computer platforms has to be performed in two steps: transferring the programs to a non-obsolete environment and executing the programs in a different rendering environment.

In [7] we demonstrated the documentation of the output formats of an early home computer system (the Philips Videopac G7400 utilizing an extension that allows the system to execute BASIC software and store and retrieve data from and to cassette tapes). We showed that even for comparatively simple systems a lot of steps are necessary to reengineer the data formats. In a case study shown in the same paper, we transferred data from various old tapes to a non-obsolete environment using a tool we developed. The data was then migrated to non-obsolete formats using signal processing techniques to convert the analog sound signal to binary data. While static data like images can then be opened in current viewers, software in BASIC source code format converted to readable text can not be executed in a current environment without further preservation actions.

Source code is one of the significant properties of software that allow us to migrate the software for preservation purposes [12]. For interpreted program languages like BASIC (compared to program languages where source code is compiled to executable software) the source code is equal to the executable software given the availability of a suitable interpreter.

Diessen et. al. describe in [18] the view-path as „a full set of functionality for rendering the information contained in a digital object". The view-path contains the hardware and all other secondary digital objects needed to render an object and also to run a certain piece of software. As an example, to run a simple JAVA program printing 'Hello World' on screen, a JAVA virtual machine, different libraries, an operating system running the virtual machine and the hardware to execute the operating system are needed. In OAIS [9] terminology the view-path contains the Access Software used to render the digital object as part of the representation information and all secondary digital objects needed to execute the Access Software.

Different strategies for preserving digital objects exist, the major ones being migration and emulation. Migration, which involves altering the original format of the digital object ([11]), is the main strategy for preserving static content. In [14] Rothenberg explains that the emulation of the logical behavior of a computer system should be sufficient on a relatively abstract level. Lorie differentiates between the archiving data and archiving program behavior. While the first can be done without emulation, Lorie argues that it cannot be avoided for the latter [10].

Execution in an emulation environment necessitates expert knowledge about utilization of the original environment and creates issues like data exchange between the emulation environment and the host environment [13]. Although the second issue was partially solved in the emulator Dioscuri, created specifically for digital preservation [16], it is still far from being a standard in current emulators.

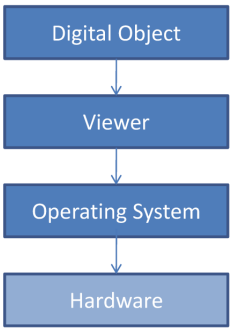


Figure 1: view-path for a generic system.

The European research project KEEP[1] performs research in legal aspects of emulation as well as develops a common platform for emulators (*Emulation Virtual Machine*) to „Keep Emulation Environments Portable". Some of the legal issues raised by KEEP also apply to the development of the emulator in this paper.

In [4] examples for the fragility of performance works based on electronics under the aspect of re-performance are provided and the question is raised, how to guarantee authenticity when preserving the electronic material. Comparing renderings of the same digital objects in different environments is usually done manually by a human observer. A case study to compare different approaches to preserve video games, with one of the approaches being emulation, was reported in [6] on a human-observable and thus to some extent subjective level. In [8] we presented case studies of interactive objects comparing the rendering outcomes of different rendering environments.

In this paper we show how the concept of a view-path can be applied to an obsolete system. We explain how software for the system is preserved using emulation by implementing an appropriate emulator. Digital preservation in mind we discuss the design decisions that have to be taken and we show discuss how the emulation results can be compared.

3. PROGRAM EXECUTION ON THE ORIGINAL SYSTEM

For identifying the elements needed for the execution of software on the original system, we first have to determine the view-path of the software.

In the most simple case the view-path of a digital object contains the digital object, the viewer used to render the object, the operating system to execute the viewer and the hardware to run the operating system as shown in Figure

[1]http://www.keep-project.eu/

Figure 2: Philips Videopac+ G7400 with plugged in Philips C7420 Home Computer cartridge.

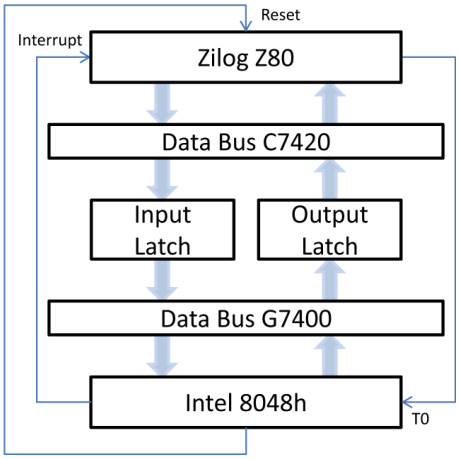


Figure 4: Communication flow between G7400 system and C7420 cartridge.

1. Depending on the digital object and the system used, some elements in the view-path can be missing. E.g. if the digital object is software, then usually the software is running directly „on top" of the operating system. In the case of early computers, the software runs directly on the hardware without the use of an operating system.

To determine the view path on the original system, information about the hardware and the software running (e.g. BIOS) has to be collected. This information can be collected using different sources like the original circuit diagrams of the system and the cartridge, disassembled code of the Z80 BIOS and the terminal software, and last but not least valuable information found out by other members of a community still working actively with the original system (expert knowledge).

The original system used to execute the digital objects is a Philips Videopac+ G7400 video game system, which is expanded to a home computer using the Philips C7420 Home Computer cartridge (Figure 2). Details about the history of the system can be found in [7]. Using the C7420 cartridge, the video game system was extended by an extra processor (Zilog Z80), more memory (RAM) and an extra operating system (ROM) implementing the programming language Microsoft BASIC-80[2]. Figure 3 shows a block diagram of important parts of both the C7420 cartridge and the G7400 System.

The communication of the C7420 cartridge with the G7400 main system is done using a program running on the Intel 8048h processor inside the G7400 that serves as a terminal program by checking the system hardware for input (keyboard and joysticks) and also issues the commands for output sent from the C7420 cartridge to the relevant registers of the Intel 8245 VDC (Video Display Control) chip and the Thomson Semiconducteurs EF9340/EF9341 chip pair inside the G7400. These 3 chips produce all the visible and audible output of the system. Communication between the software running on the Z80 processor and the software running on the 8048h processor is managed by using two 8-bit registers that serve as a read and write latch. The Z80 processor writes information to the latch and then sets an input line on the 8048h processor. By checking the input line, the 8048h knows if information is available and proceeds reading the latch. For the other direction the 8048h writes to a different latch and sets a line that is connected to the Interrupt line of the Z80 processor, thus triggering an interrupt service routine on the Z80 that then can read the latch. Additionally the 8048h can send a RESET signal to the Z80 to reset the processor. The communication flow can be seen in Figure 4.

The BIOS, which is run on the Z80 processor, executes BASIC commands either entered by the user or stored as a program with line numbers. Results of operations are sent to the relevant registers on the G7400 using the described flow of communication. Commands accepting input are receiving the relevant input data from the G7400. Additionally to the data exchange with the G7400, the C7420 can store and retrieve data from an audio source connected directly to the cartridge using microphone / headphone plugs.

The resulting view-path for the G7400 system with C7420 cartridge can be seen in Figure 5. The digital object, in this case a BASIC program, is executed by the BASIC interpreter of the operating system. The BASIC interpreter is run on the Z80 CPU. Additionally, in this case a second branch of the view-path exists, which handles the input and output. In parallel to the operating system running on the Z80 processor, a terminal program for communication with the Z80 is run on the 8048h CPU, communicating input and output data between the G7400 system and the C7420 cartridge.

4. IMPLEMENTING THE VIEW-PATH IN AN EMULATOR

As we did not want to start working on the G7400 and C7420 emulator from scratch, the existing open source emulator

[2]Microsoft BASIC - Wikipedia: `http://en.wikipedia.org/wiki/Microsoft_BASIC`

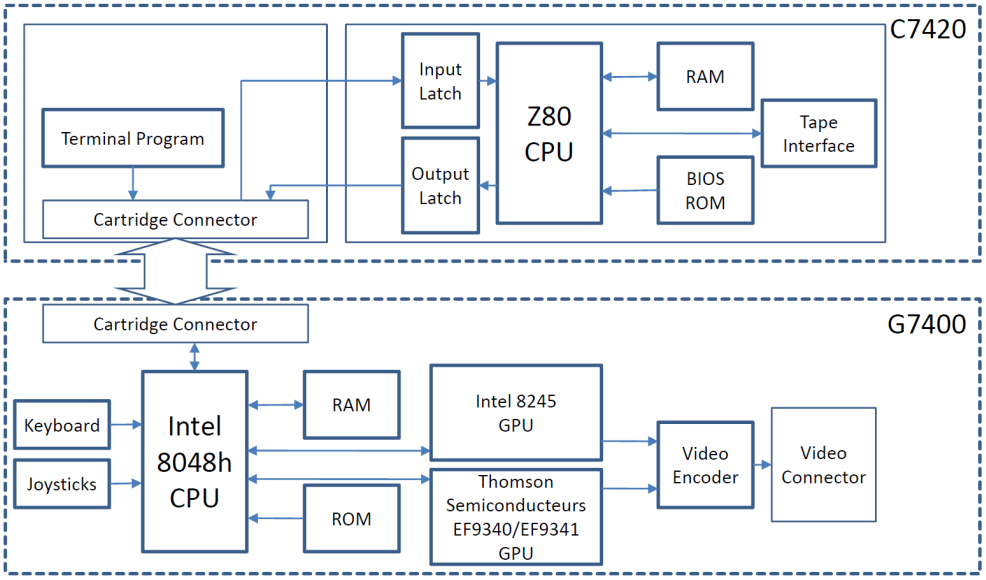


Figure 3: Block diagram of C7420 Home Computer cartridge and Philips Videopac+ G7400 system. Connection between cartridge and system is done using the cartridge connector. CPU - Central Processing Unit, GPU - Graphics Processing Unit, RAM - Random Access Memory, ROM - Read Only Memory.

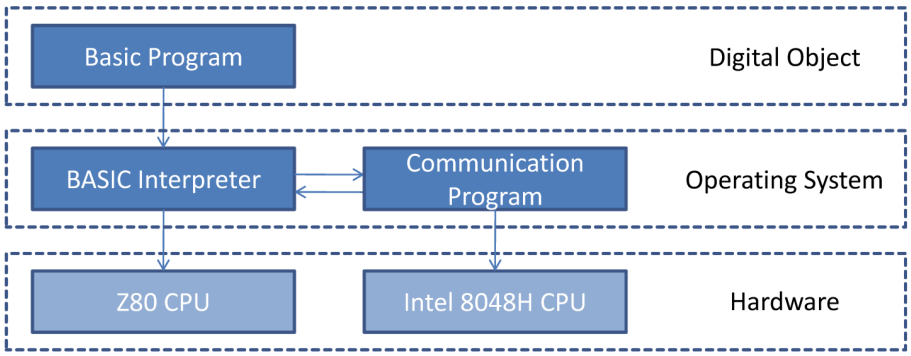


Figure 5: view-path for program execution on G7400+C7420.

O2EM[3] was used as a starting point. O2EM initially was written in 1997 as an emulator for the video game system Magnavox Odyssey2, which is the American version of the Philips Videopac G7000. It was later modified for supporting the different screen timing of the European system as well as the additional functionality of the successor of the Philips G7000, the G7400. The emulator is written in the programming language C, and is thus portable to different systems without changes.

To integrate C7420 emulation into O2EM we first have to integrate emulation for the Z80 processor that would run side by side to the original 8048h emulation. An existing emulator of the Zilog Z80 [4] programmed by Marat Fayzullin is used. Using a separate module for emulating the Z80 processor component also follows the principle of modular emulation as described by van der Hoeven et. al. in [17]. By using a Z80 processor emulation that is already proven to work in other emulators we can make sure, that the development effort on our side is reduced, minimizing also the risk of introducing erroneous emulation behavior by relying on existing, tested modules. Integration of the processor emulation consists basically of the following steps:

Z80 Memory Access and Interrupt After defining the 64 KByte memory of the C7420 as an array, the BIOS for the C7420 is loaded into the first 8 KBytes of the memory. Function prototypes provided by the Z80 emulator to access the memory are filled with code to access the memory (fetching instructions from the memory and reading and writing data). The prototype function checking for interrupts has to be adapted to signal an interrupt to the Z80 if the 8048h emulation sets the corresponding variable.

[3]O2EM - Sourceforge: `http://o2em.sourceforge.net/`
[4]Marat Fayzullin Emulation Resources: `http://fms.komkon.org/EMUL8/`

Z80 Input and Output Functions The Z80 processor has instructions for writing to output ports and also reading from them. These ports are used to access the latches for communication of the Z80 processor with the 8048h processor. The prototype functions are implemented to read from the latch defined at port 0xC0 and write to the latch defined at port 0xE0, as well as setting the T0 line of the 8048h.

I8048h Instructions, Input and Output Functions The 8048h instructions to check T0 line were previously only implemented to support a different kind of expansion for the G7400 system. These instructions have to be adapted in order to read the line that is set by the Z80 processor and reset it (to tell the Z80 processor that the 8048h recognized a written byte). Reading and writing to external memory also has to be adapted to read from the latch-register defined as external memory on address 0xE0 and write to the latch register defined as external memory on address 0xC0. Additionally, the write-function to the output ports of the 8048h has to be adapted, as pulling the lower two bits of Port 1 to low is supposed to reset the Z80 and pulling just Bit 1 of Port 1 to low signals an interrupt on the Z80.

Execution of Z80 cycles Finally the emulation main loop has to be extended to include the execution of Z80 instructions. The 8048h processor is running at a clock rate of 0.394 MHz internally, while the Z80 processor is running at a 3.547 MHz clock rate, which makes it roughly execute 10 clock cycles for every 8048h clock cycle. Completely accurate cycle exact timing was not a necessity, as the communication between Z80 and 8048h is based on a handshake protocol, so one waits until the other provides the necessary data. The main execution loop sets the counter of cycles to execute to 10 and invokes the Z80 emulation.

To actually synchronize the emulation of the 8048h and the Z80 and implement the aforementioned steps, debug output of instructions of both processors is enabled and the log analyzed to find out exactly, which processor is doing what at a given point in time. By debugging through the assembler instructions of both processors, the handshaking can be established and the emulator starts up with the start screen of the C7420 Home Computer cartridge as shown in Figure 6.

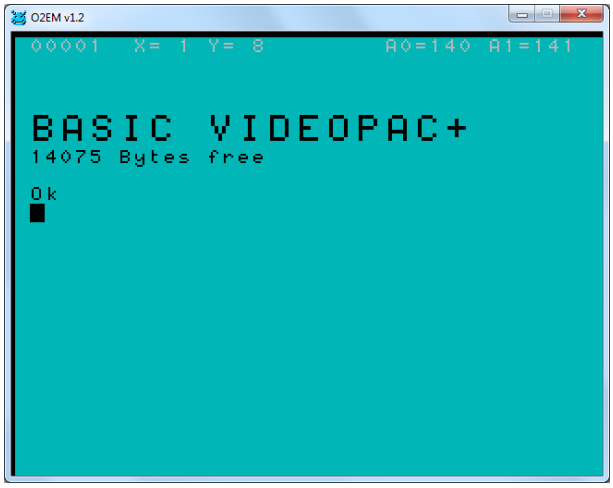


Figure 6: Start screen of C7420 Home Computer cartridge on O2EM emulator.

5. DATA INJECTION

After establishing the emulation of C7420 Home Computer cartridge, the next step is to enter data into to the emulated environment. Three options for data input are available on the original system. Below we describe these three options and the challenges they present for emulation.

5.1 Keyboard

An obvious method of data entry to the emulated environment is a key press. The previous implementation of the keyboard routine mapped every key on the original G7000 system keyboard to a key on a standard PC keyboard. This was sufficient for the currently emulated programs as the extra keys of the G7400 keyboard were not used in any of the supported programs.

In a first step we correct the keyboard routine to support the extra two rows of keys on the G7400's keyboard. This provides us with the possibility of mapping every key on the G7400 keyboard to a key on a modern keyboard. Unfortunately, the differences between current keyboards and the original G7400 keyboard are quite significant. As an example, a special key providing opening and closing brackets ('[' and ']') exists which is not directly to be found on a modern keyboard but only reached through key combinations. Additionally, various key combinations create different effects, for example the number sign ('#') is printed on the G7400 keyboard as a combination of the SHIFT key and the number '0', whereas a modern keyboard has its own key for it.

The BIOS of the G7400 checks the keys by going through every line of keys on the keyboard and reporting which key is pressed. Combinations of keys (e.g. SHIFT and a number) are recognized in the terminal software of the C7420 running on the 8048h processor. This software converts the pressed key to an ASCII encoded character depending on the combination of keys pressed and sends the ASCII code to the Z80 BIOS routine.

To improve the keyboard routine, we identify the following levels where it can be intercepted:

Z80 BIOS Directly inserting key-presses into the keyboard routine of the Z80. The Z80 reads the keys received from the terminal program running on the 8048h and writes them in a keyboard buffer. Keys read in ASCII-format from the host-keyboard can be directly written into the keyboard buffer (with the exception of characters that have a different code on the C7420 system). This would be a special routine only working for the C7420 BIOS, as it uses specifics otherwise not found on the system. It also would not be compatible with

the current keyboard routine.

Communication interface Alternatively, keys can be written to the memory of the 8048h. As the keyboard routine in the terminal software already converts the key presses to ASCII, keys could be written as received from the keyboard functions. This method like the previous one would be a special implementation for the C7420. The existing hardware emulation would have to be disabled to not interfere with the other routine.

Hardware level Adapting the keyboard routine on the hardware emulation level offers the most compatibility not only for the C7420 Home Computer cartridge but for all other software developed for the G7400 system as well. Instead of the current implementation to have a one-to-one relationship between a key on the host keyboard and a key on the emulated hardware, with the flaws described above, a new routine could do a mapping of the actually entered character on the host system and set the appropriate keys in the emulated environment to simulate key-presses corresponding to the entered character.

We decided to extend the keyboard routine on the hardware level to reach the best compatibility for all programs running on the hardware. In a first step we create a mapping for all useful key-presses on the G7400 (e.g. combinations like 'CONTROL', 'SHIFT' and a character don't have any effect on the C7420, and even thought they could be theoretically read by replacing the G7400 BIOS routines by a self-written routine, the ergonomics of the membrane keyboard make it hard to press two keys at the same time). Next we replace the routine that reads the state of the mapped keys by a routine that first reads the ASCII Code of the entered character (considering modifier keys like Shift or Control), and sets the corresponding keys on the G7400 emulation using a „best guess" strategy to decide what the user actually wanted (e.g. entering '=' sign on the host keyboard (using a combination of different keys on the host keyboard) is mapped to pressing the '=' key on the G7400 keyboard. Likewise entering ';' on the host keyboard emulates a key press of the Clear key and the Shift key on the G7400 keyboard, which - in the original system - produced the semi-colon. Some of the keys had to be emulated by non-obvious combinations, for example one key for creating a character consisting of two dots, not available in ASCII or an modern keyboard, was simulated by entering '§'.

To test the validity of the keyboard routine, we wrote an assembler routine that reads out the pressed key and compares the results of the program on the real hardware and the emulator. Entering key-presses to the emulated C7420 environment also now creates the expected results. We also checked some samples of other software running on the emulator to make sure that the new keyboard routine did not break other software for the system.

5.2 Joysticks

The original system has two joysticks that are emulated by O2EM either using actual joysticks connected to the host environment or keyboard emulation for the joysticks. The polled data is provided to the emulated environment as soon as the BIOS of the G7400 tries to read the hardware ports. It is then handed over to the BIOS running on C7420 and can be read using the correspondent BASIC commands (e.g. STICK(0)). As the joysticks were already properly emulated by the original emulator, no additional actions had to be performed.

5.3 Files

Besides data injection through control devices, the C7420 supports the loading of files from an audio signal connected through a microphone jack. In this section we will show different possibilities of loading a file into memory.

Hardware Emulation On a hardware emulation level, the component for reading data from the audio source, converting it to a digital signal and providing it on the input port of the Z80 is the most complex one. Basically, when the user tries to load a file using the 'CLOAD' command, the bits provided in the audio stream are decoded, assembled to a byte and written to the appropriate memory location. By reengineering the original BIOS routine of the 'CLOAD' command and based on the format as described in [7] we were able to create a routine that emulates that behavior of the original tape interface and provides the correct data in the correct timing to the CPU. The original tape was simulated by providing a directory in which the different files are stored. Using 'CLOAD' without a filename loads the file first written into the directory, subsequent calls of 'CLOAD' load the next file respectively. Using 'CLOAD' with a filename loads the file with the specified filename. 'CLOAD' supports loading of every file type supported by the C7420, i.e. BASIC programs, screenshots, data, and memory dumps.

Direct Writing to Memory An alternative to the aforementioned method of hardware emulation is to load a file into memory and directly write the loaded bytes into the correct memory locations. For this purpose the behavior of the original 'CLOAD' has to be reengineered even more to find out what all memory positions are affected (e.g. counter for free memory). Using this method we implement a special key that presents the user with a file-browser-dialog to select a file. Only BASIC programs can be stored using the direct memory method.

Both of the aforementioned methods result in the same memory structure when loading a file, with writing directly into memory being much faster (as the file is instantly loaded) whereas the hardware emulation preserves the original timing and thus needs a few minutes for programs with more than 100 lines. Using the hardware emulation it is possible to have programs load and save data from within using the original BIOS functions.

The data loaded from the tape interface is basically in the exact same format as written into memory (with the addition of leading and trailing bytes and some start- and stop-bits to separate bytes). To provide better support for using the emulator as a cross-programming-tool, we also implement implicit migration of BASIC files in text format. Loading

a text file containing human readable BASIC source code is automatically detected and migrated back to the original binary format with encoded line numbers and encoded BASIC commands, so it can be used again in the original environment, the C7420.

6. DATA EXTRACTION

While data injection is an important issue to execute and interact with software in the emulated environment, for some digital preservation applications it is necessary to extract data from the emulated environment. Especially if emulation is used to access data stored in its original format and the data has to be used in the host environment, methods of copying data to one's current environment have to be provided. The methods for data extraction we implemented in the emulator are listed below.

6.1 Files

Using an emulator to modify data stored in an obsolete format makes it necessary to be able to save previously loaded files again. Again, two different methods are implemented:

Hardware Emulation The BASIC command 'CSAVE' for saving data is implemented analogue to the command for loading files. We again have to reengineer the format by examining the code of the BIOS written in Z80 machine language to observe, what data is written to the output interface. The data stored by the BIOS is written to an array and saved under the filename given with the command. 'CSAVE' works for all possible variations, saving programs, data, screenshots and memory dumps.

Direct Read From Memory As with 'CLOAD' a function to directly write a BASIC program to disk is provided. As the format of storing BASIC programs in the memory of the C7420 was analyzed for creating the other file functions, it was also possible to create a function to provide a dialog to the user to ask for a filename and directly dump the memory in the correct format to a file.

As with 'CLOAD' the resulting file is the same in both cases, with the hardware emulation being compatible to all formats and the direct read from memory version being easier to use without expert knowledge and being considerably faster. The choice of type of BASIC file (either in text format for easy readability or in binary format as originally created by the system) can be specified as a command line option for the emulator.

6.2 Clipboard

One feature hardly present in emulators today but crucial for their use for digital preservation purposes is the possibility to extract rendered text in machine-readable form as separated characters from the emulated environment for use in the host environment. As the original environment in the C7420 does not support marking regions of text on the screen, and putting it in an internal clipboard, we decided to implement a function that copies the whole screen content as characters into the clipboard of the host system, so the text can be pasted into any application. Two different hook points for extracting data from the C7420 are possible:

Extraction from C7420 screen buffer The C7420 Home Computer cartridge holds an internal representation of the screen buffer for manipulation through the Z80 in the Z80 memory area (RAM). Extracting the characters from there would be possible by reengineering the memory location the screen data is saved at, as well as the format it is saved in. This would be the preferred option if the data was not rendered in the hardware chip as text on the screen.

Extraction from emulator screen buffer The G7400 uses a teletext type of display chip for rendering graphics of the C7420. Thus a representation of the screen data (the characters) has to be held in the video screen buffer for rendering the image. By extracting data from the video screen buffer we not only create the possibility of copying data from the C7420 cartridge but also from all other software for the G7400 using the video chip.

We decided to go with the more generic version and extract the data directly from the video memory of the emulator. Depending on the operating system different routines for copying data to the clipboard has to be implemented. The data that is extracted is in ASCII, so we can directly use it for copying it to the clipboard. The video chip is able to apply certain special effects on the characters (e.g. double size, blinking characters, underlined characters). As we need to get a text representation of the data for later usage in other applications we decided to ignore the format and just copy the actual characters to the clipboard. As not all the characters have the same code representation as in a current ASCII format table, a conversion for certain characters is performed while copying the data.

6.3 Screenshots

Screenshots of the emulated environment can be used e.g. to compare emulation results with the original environment. Extracting data in the form of screenshots can be done using one of three different methods on different levels of the emulation:

In the Emulated Environment Using the screenshot feature of the C7420 (the 'CSAVES' BASIC command) the screenshot can be saved to a file and converted to a non-obsolete format using the tool we developed in [7]. Using this method it is possible to compare the principal rendering inside the emulation environment. It can not be checked if the emulator renders the image correctly on the host system.

Inside the Emulator The emulator O2EM has a built-in feature that allows saving screenshots of the rendered environment. Using this feature it is possible to manually save screenshots at certain points in the emulation.

From the Host Environment Using a screenshot tool inside the host environment automatic screenshots at different time points can be taken as well as a video of the emulation.

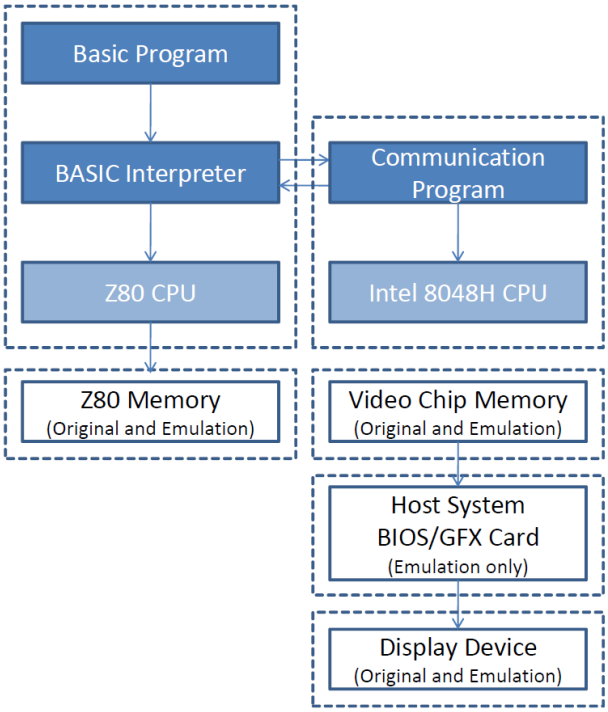


Figure 7: Different renderings in the view path of the C7420 Home Computer cartridge.

The resulting screenshots can be used e.g. to compare rendering results of different rendering environments for preservation planning purposes as described in [8].

7. EVALUATING RENDERING RESULTS

To select the best preservation solution for a certain scenario, it is necessary to compare all available preservation actions. In [2] Becker et.al. describe a preservation planning approach based on comparing significant properties of digital objects before and after applying a preservation action. While on migrated data the digital object before and after migration can be compared, the task is different when dealing with emulation. Instead of comparing the digital object, renderings of the digital object in different rendering environments are compared.

Results of rendering can be compared on different levels. Figure 7 shows the different levels on which an image is rendered inside the view-path of the C7420 Home Computer cartridge in conjunction with the G7400 system.

In detail the levels on which we can compare the rendering results are:

Z80 Memory The BIOS running on the Z80 has an internal representation of the screen memory that can be extracted using the screenshot feature 'CSAVES'. Doing this on the original system and on the emulated system, we receive two files which can directly be compared. If the files are identical, then the emulation of the Z80 CPU is correct (for the rendering of the test digital object). Yet, we cannot ascertain, that the actual rendering as provided by the emulator matches the rendering of the original system.

Video Chip Memory Another representation of the rendered object exists in the Memory of the video chip. This memory region is emulated in the emulator and can be read out. Unfortunately it cannot be read on the original system without directly reading the signals from the hardware and decoding them accordingly.

Host System BIOS The emulator renders the image stored in the video chip registers. The image is rendered and saved either in the Host system representation of the screen content or directly in the video card memory. Obviously this representation of the rendering exists only in the emulated rendering environment. Using this representation (basically creating a screenshot of the emulator's output) we can compare different rendering environments running on a host system (e.g. emulator of architecture level, high level emulator). In [8] we demonstrate how the rendering results of different rendering environments can be compared by using the characterization language XCL as described in [3] for objectively comparing the significant properties of two screenshots.

Display Device Finally, a comparison on the level of the display device (comparing the output of the original system on a display device with the output of the emulator on a different or even the same output device) can be performed. This comparison is usually done manually and subjectively by the human preservation planner.

Not only the level of extraction of an image for comparison is relevant, also the time line is important. Usually, especially with interactive and dynamic software, we are not only interested in a screenshot at a certain point in time, but either a series of screenshots or a continuous extraction of a video stream, which also allows the comparison of factors like timeliness and synchronicity, e.g. with sound output, compared to the original.

While the emulator supports already the extraction of screenshots (activated by pressing a key), a continuous extraction of images or extraction of images after a certain amount of elapsed time or executed machine cycles is currently not supported.

8. OTHER PRESERVATION ACTIONS

Executing programs using emulation on a hardware level is only one of the different alternatives that can be used for preserving software. Figure 8 shows the different levels in the execution view-path of the C7420 and also lists preservation action strategies for each of the levels.

8.1 Hardware Level

On the hardware level the emulator that was implemented can be used to preserve the system's behavior and thus create a rendering environment where the original operating system software (BIOS) can be used to execute the programs. As shown before, the reengineering effort necessary

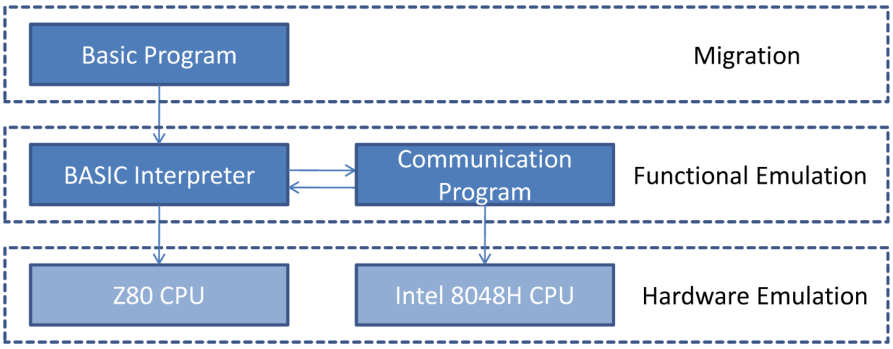


Figure 8: Preservation actions for different layers of view-path.

to implement an emulator is quite high, even though this method is probably the most accurate one.

8.2 Functional Level

Creating an emulator for the BASIC-programs not on a hardware level but on a functional level would require to implement an interpreter for the BASIC-code, that emulates the functions of the original BASIC-commands. Instead of executing the underlying Z80 machine language code in the BIOS if e.g. a „PRINT" command is executed, the interpreter would emulate the behavior of the command, i.e. printing characters on the screen. Data extraction and injection is obviously much less complex, as the rendering environment can be directly manipulated and the behavior of each command can be controlled.

8.3 Source Code Migration

A completely different strategy than emulating the system on a hardware level or emulating the commands on a functional level is the migration of the BASIC-programs to a non-obsolete programming language. Running a parser over the programs and migrating every command to a representation in a non-obsolete programming language allows us to create stand-alone versions of the programs that can be run without the need of an emulator program. While some of the commands would be quite easy to migrate (e.g. mathematical operations), others would involve more complex implementations (e.g. setting a different screen mode, displaying characters on the screen). Another obstacle to overcome in the special case of the C7420 is the flow of program execution, if the target language is a structured programming language instead of an unstructured one that is line-based like the used Microsoft BASIC-80 language. Jumps in the program between line numbers (and even to calculated line numbers stored in variables) have to be converted to different types of control flow statements (e.g. loops or choices). The principal possibility of this conversion has already been shown in [1].

9. CONCLUSIONS AND FUTURE WORK

In this paper we described how an emulator for an early home computer system was developed. We presented the reengineering work involved in enabling emulation of the system itself as well as reengineering necessary for emulating save and load functions. The emulation was implemented keeping digital preservation applications in mind, so data injection and extraction with ease of use for users without expert knowledge of the system was implemented. We described what challenges arose while implementing the emulation and what design decisions were taken and why. We also explained how we were trying to keep special digital preservation requirements in mind when implementing certain features like extracting data from the emulation environment. We showed how different rendering environments can be compared and on what levels specifically for the machine in the case study, and how this either is already supported or would have to be implemented in the future. Finally, we discussed other options for preserving software for the home computer system evaluated like source code migration and high level emulation in the form of a BASIC interpreter.

The work performed for this emulator shows how complex the task to develop an emulator is and what steps are involved especially for a system without proper and open documentation. It further shows what design decisions arise during the development of an emulator especially when having a long term approach in mind and not only a short term solution for executing software of a recently obsolete system.

The implementation of the emulator was considered a success as the digital objects migrated previously from audio tapes could be injected and successfully executed in the emulated environment. The case study also showed that the actual implementation of the emulation of the C7420 Home Computer cartridge was in this special case a comparatively less complex task, as a well documented and already emulated Z80 processor was used as the central processing unit of the C7420. The more time intensive task was the reengineering of the components used for data injection and data extraction, on one hand the emulation of the C7420 tape interface, and on the other hand the proper emulation of keyboard input and data extraction to the clipboard.

One important lesson learned while implementing the emulator was that the input and output routines will most likely have to be adapted at the time of dissemination of archived data. A change in layout of keyboards used between archiving the emulator and the data to be rendered will already enforce a change in the keyboard routines of the emulator.

If the method of entering data changes from keyboard to something else (which is not an unlikely scenario given a time frame of 50 to 100 years) the mapping of data input has to be completely adapted. Similarly, the data extraction from the emulated environment in the shown example already enforced a change in certain character codes. Given a longer time frame between archival and reuse of the archived emulator, these kind of adaptions are even more likely to be necessary, even if the environment for the emulator (e.g. an emulation virtual machine as described in [15]) keeps the emulator executable.

For future work we plan to implement other strategies for preserving the C7420 software as listed in Section 8. A comparison of the different strategies on different levels of the view path will be performed to show how the quality of emulation can be objectively measured. The results of the work carried out on the fairly simple C7420 Home Computer cartridge system will then be applied to more complex systems.

10. ACKNOWLEDGMENTS

The research was co-funded by COMET K1, FFG - Austrian Research Promotion Agency and by European Community under the IST Programme of the 7th FP for RTD - Project ICT-269940/TIMBUS.

11. REFERENCES

[1] E. Ashcroft and Z. Manna. *The translation of 'go to' programs to 'while' programs*, pages 49–61. Yourdon Press, Upper Saddle River, NJ, USA, 1979.

[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, A. Rauber, V. Heydegger, J. Schnasse, and M. Thaller. Systematic characterisation of objects in digital preservation: The extensible characterisation languages. *Journal of Universal Computer Science*, 14(18):2936–2952, 2008. `http://www.jucs.org/jucs_14_18/systematic_characterisation_of_objects`.

[4] A. Bonardi and J. Barthélemy. The preservation, emulation, migration, and virtualization of live electronics for performing arts: An overview of musical and technical issues. *J. Comput. Cult. Herit.*, 1(1):1–16, 2008.

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

[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, M. Ghete, A. John, C. Lederer, and A. Rauber. Migrating home computer audio waveforms to digital objects: A case study on digital archaeology. *International Journal of Digital Curation (IJDC)*, 6(1):79–98, 2011.

[8] 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.

[9] ISO. *Space data and information transfer systems – Open archival information system – Reference model (ISO 14721:2003)*, 2003.

[10] R. Lorie. A project on preservation of digital data. *RLG DigiNews*, Vol. 5 (3), 2001. `http://www.rlg.org/preserv/diginews/diginews5-3.html#feature2`.

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

[12] B. Matthews, B. McIlwrath, D. Giaretta, and E. Conway. The significant properties of software: A study. JISC Study, 2008. `http://www.jisc.ac.uk/media/documents/programmes/preservation/spsoftware_report_redacted.pdf`.

[13] T. A. Phelps and P. Watry. A no-compromises architecture for digital document preservation. In *Proceedings from 9th European Conference on Research and Advanced Technology for Digital Libraries*, pages 266–277, 2005.

[14] J. Rothenberg. *Using Emulation to Preserve Digital Documents, Tech. Rep.* Koninklijke Bibliotheek, 2000.

[15] J. Slats. Emulation: Context and current status. Tech. Rep., 2003. `http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf`.

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

[17] J. van der Hoeven and H. van Wijngaarden. Modular emulation as a long-term preservation strategy for digital objects. In *5th International Web Archiving Workshop (IWAW05)*, 2005.

[18] R. J. van Diessen. Preservation requirements in a deposit system. *IBM/KB Long-Term Preservation Study Report Series Number 3 Chapter 3*, 2002. `http://www-05.ibm.com/nl/dias/resource/preservation.pdf`.

Developing Virtual CD-ROM Collections: The Voyager Company Publications

Geoffrey Brown
Indiana University School of Informatics and Computing

ABSTRACT

Over the past 20 years, many thousands of CD-ROM titles were published; many of these have lasting cultural significance, yet present a difficult challenge for libraries due to obsolescence of the supporting software and hardware, and the consequent decline in the technical knowledge required to support them. The current trend appears to be one of abandonment – for example, the Indiana University Libraries no longer maintain machines capable of accessing early CD-ROM titles.

In previous work we proposed an access model based upon networked "virtual collections" of CD-ROMs which can enable consortia of libraries to pool the technical expertise necessary to provide continued access to such materials for a geographically sparse base of patrons who may have limited technical knowledge.

In this paper we extend this idea to CD-ROMs designed to operate on "classic" Macintosh systems with an extensive case study – the catalog of the Voyager Company publications which was the first major innovator in interactive CD-ROMs. The work described includes emulator extensions to support obsolete CD formats and to enable networked access to the virtual collection.

Keywords

emulation,digital preservation,voyager company

1. INTRODUCTION

Emulation has been widely discussed as a preservation strategy for digital artifacts such as multimedia presentations that are intimately tied to their original hardware/software platform for interpretation [17, 9, 12, 23, 25, 24]. Emulation has been successfully tested to preserve individual artifacts such as the BBC Doomsday book project, various multimedia art works [18, 30] and is widely used for the preservation of console games [11]. At this point it is clear that emulation can be used to successfully access software on many obsolete platforms. The fundamental question addressed by this paper is how emulation technologies might be scaled to support convenient access to large collections of born-digital materials.

We have previously proposed a general model for preserving "virtual CD-ROM" collections and explored the use of emulation of Windows based platforms [33]. In this paper, we extend this work to emulation of classic Macintoshes through a significant case study – the CD-ROMs published by the Voyager Company. Although the work required non-trivial modifications to an existing open-source emulator, we successfully demonstrate that the fundamental model is both sound and practical for those CD-ROMs that depend upon classic Macintosh environments. This work also explores an important architectural alternative to our previous work. Previously, we provided custom compute environments by using artifact specific customization on a client-side "generic" emulation environment. In this work we utilize pre-customized server-side emulation environments accessed from the client machine. Both approaches have clear advantages and it is gratifying to demonstrate that the latter also works well in practice.

The Voyager Company led by Bob (Robert) Stein is widely viewed as one of the first and most influential publishers of interactive media on CD-ROM [31, 35, 36]. Over the period 1989-1997 the company published approximately 75 CD-ROMs as well as a large number of "extended" books – the latter are the natural ancestors for today's Kindle and other electronic books. Unfortunately, this pioneering work is largely inaccessible to contemporary users and scholars because of its dependence upon obsolete versions of the Macintosh operating system and related software. Indeed, it was a chance meeting with John Eakin, a Professor of English at Indiana University, whose scholarly interest in biography and in particular "The Complete Maus: A Survivor's Tale" [CD-29][1] was the catalyst for the work described in this paper. The Voyager version of Spiegelman's Pulitzer Prize winning work includes the original book augmented through hyperlinks with interviews, videos, as well as preliminary drawings in a delightful interactive CD-ROM. As Professor Eakin noted, while the Indiana University Libraries hold a copy of this work, it has an attached sticker announcing that "We [the Libraries] no longer have systems to support format" [7].

[1]Labels of the form [CD-i] refer to the table in AppendixA

iPRES2011, Nov. 1–4, 2011, Singapore.

The basic tools necessary to execute (most of) the Voyager CD-ROMs in a software emulation of the classic Macintosh have existed for several years in the form of two open-source (and tightly interwoven) projects – SheepShaver [28] and BasiliskII [3]. SheepShaver emulates the later PowerPC based Macintoshes while BasiliskII emulates the earlier 680xx based Macintoshes. Both are capable of running versions of the Mac operating system that can support the Voyager CD-ROMs. However, there is a significant gap between "capable" and "practical" for the casual user which might represent an insurmountable barrier for the less technically sophisticated. For example, installing an emulation system capable of executing the Voyager CD-ROMs requires installing (and possibly compiling) the emulator, finding and installing suitable versions of the Mac operating system and system ROM, and finally installing the CD-ROM itself.

A goal of the work described in this paper is to enable easy access to the CD-ROMs for casual users and scholars by developing the technologies required to make network access to a collection of the CD-ROMs practical with minimal software installation on the end-user's workstation. In particular, we discuss an approach that encapsulates all the necessary operating system and application software in a small down-loadable package which may be executed on any machine with the necessary emulator installed and in which the CD-ROMs themselves are accessed over the Internet. The techniques presented make it possible for libraries holding a collection of CD-ROMs to provide public access through modern workstations with no technical knowledge required by patrons. Furthermore, we show how existing distributed file-system software can enable the creation of consortia to pool the resources and technical knowledge necessary to support geographically distributed patrons while enabling strict access controls.

Clearly, this work raises serious legal questions. All of the Voyager CD-ROMs as well as the Macintosh operating system (and system ROMs) are protected by copyright and it is certainly not legal to provide unconstrained access to these materials. We assume that the best case scenario will require clear controls limiting access to authorized patrons and the technologies described in this paper enable such access controls.

The preservation challenges presented by the Voyager CD-ROMs were well described by Jeff Martin [16]. He provides a clear, basic overview of the history of these publications, interesting observations based upon interviews with Voyager programmers about the underlying software, and the results of testing several (4) of these CD-ROMs under OS 8.5.1 on a legacy Mac as well as under the "classic mode" which was part of the early OS X distributions. Most (3) of these examples performed poorly even under 8.5.1 suggesting that support for the Voyager CD-ROMs requires access to earlier versions of the Mac operating system. Martin did not explore emulation as a possible approach to preservation, his testing was limited to 4 CD-ROMs versus 48 tested for this paper and he provided no solutions to the 75% failure rate he encountered under OS 8.5.1. In contrast, the methodology described in this paper was successfully used on all 48 CD-ROMs tested (more than half of the Voyager publications), although, as we note, there are a handful of titles whose behavior is "fragile."

The remainder of this paper is organized as follows. In Section 2, we describe in greater detail our vision for virtual CD-ROM collections and discuss how existing technology can form a viable foundation for creating such collections. We provide a brief overview of the Voyager CD-ROMs with an emphasis on those tested for this work in Section 3. We discuss the key emulation and CD-ROM imaging technologies in Sections 4 and 5 and conclude with a discussion of copyright in Section 6 and results in Section 7.

As mentioned above, existing emulation tools can support most of the Voyager CD-ROMs with the notable exception of those utilizing mixed-mode data/audio CD formats. However these happen to be some of the most novel of the Voyager publications including their first – a hyper-media presentation of Beethoven's Ninth Symphony – as well as Pedro Meyer's "I Photograph to Remember" [CD-2] which was the first CD-ROM with continuous sound and images ever produced [19]. As part of this work, we have extended both SheepShaver and BasiliskII to support these CD-ROMs and describe the necessary changes in Section 4.

2. VIRTUAL CD-ROM COLLECTIONS

Although the Voyager CD-ROMs have substantial historical significance, they, and most other published CD-ROMs, are destined to have a dwindling user base whose expertise in the systems required to use them is in sharp decline. The physical machines required to execute them have already disappeared from most educational institutions – even the operating systems are increasingly hard to find; at Indiana University, which once had many hundreds of "classic macs", only one person within our University IT Services had distribution disks of the corresponding operating system software. The physical copies of these CD-ROMs are disappearing from library shelves – in seeking examples for this paper we made extensive use of inter-library loan and we found that many cataloged copies of Voyager CD-ROMs are either missing or damaged.

The long-term probability for individual libraries providing physical access to the Voyager and other published CD-ROMs is nearly nil. The user base is dwindling, the existing hardware and software support disappearing, and the physical media degrading. While we believe these materials have substantial historical significance, their ultimate survival depends upon spreading the preservation burden across many institutions through a virtual collection that enables networked access for a sparsely distributed base of patrons using modern work-stations.

A virtual CD-ROM collection consists of two primary components – one or more servers that maintain bit-faithful images of the CD-ROMs and corresponding support software, and patron workstations with appropriate emulation software installed. Libraries and educational institutions could collaborate in creating images of CD-ROMs in their collections as well as customizing supporting software images for these CD-ROMs. In our previous work we assumed a client-side emulator pre-configured to execute a generic Windows XP environment and utilized a helper application to customize this environment for a particular CD-ROM.[33]

Where emulator environment size is substantial compared to CD-ROM size (3-4 times for Windows XP) this represents a substantial space savings. In the present work, we experimented with custom "Mac OS" environments stored on the server and accessed in a non-destructive manner by the client emulator. For this work the additional storage overhead is only about 20% and hence this simpler model is viable. The advantages to the simpler model are that it is more robust and provides tighter control over materials covered by copyright.

The enabling technology for this vision is a distributed file system. The basic idea is that emulation software is provided remote access to CD-ROM images through the file system. For example, the emulator "mounts" a CD-ROM by opening the corresponding networked file. The actual bits corresponding to a CD-ROM are pulled to the emulator as needed. In contrast a web-server based solution would require copying images before access. Consider that a CD-ROM may contain 650 MBytes; copying such a CD-ROM across a network before use could involve a substantial delay. Playing the audio portion of a CD-ROM requires a bandwidth of less than 1.25 Mbits/second. These bandwidth requirements can be met by most DSL connections. We expect a patron in a library would see no discernible performance penalty over local copies of the Voyager materials. Indeed, we store CD-ROM images on the university research file system in Indianapolis, yet work in Bloomington. we perceived few performance issues over an (optimistic) 3 Mbit home DSL connection and no issues over a much faster office connection. Thus, using a distributed file system to store CD-ROM images offers near instantaneous access while a web-server approach requires users to endure substantial delay. Furthermore, as discussed below, a distributed file system more clearly meets the spirit of copyright restrictions because the CD-ROM images are streamed during access and are never copied to the end-user's work-station.

There are many examples of network file-systems including NFS, Samba, WebDAV, and others. We use OpenAFS [21] (the Andrew File System); however, our work is not tied to this system. In addition to support on all the major operating systems, the major advantages of OpenAFS for virtual CD-ROM collections are:

- Federated authentication (in this case Kerberos)
- Fine-grained access control through access control lists
- Local volumes
- Unified name space

Federated authentication through Kerberos means that any educational institution (currently many) with Kerberos-based user authentication could authenticate their users on behalf of the collection. As mentioned above, libraries must be able to control patron access to specific items. OpenAFS access control to individual files can be specified at both the user and machine level. For example, a library could limit access for their images to specific machines or to specific local users. However, OpenAFS enables inter-library loan by providing a mechanism enabling temporary access to users at other institutions. By storing images on local volume stores individual libraries can satisfy any mandate on where digital copies may be kept. Finally, a unified name space means that any user can refer to a file by a single global name. For example, the files we maintain are in the (OpenAFS) directory:

```
/afs/iu.edu/home/projects/sudoc/Voyager
```

Any user anywhere with the correct access permissions can read this directory using this name on any machine supporting OpenAFS. No other distributed file-system currently provides all of these capabilities.

3. VOYAGER CD-ROMS

The Voyager Company was notable both for its pioneering CD-ROM titles as well as for their broad range. The company produced a diverse range of content including music, movies, books, poetry, and art. Most of their titles are based upon a simple premise – migrate an existing work in another medium to CD-ROM while expanding the original content with audio, video, and background material. In the best of these "migrations", the unique (pre-web) ability of CD-ROM content to rapidly and randomly access audio and video provided a vastly enhanced experience. For example the audio recordings developed by Robert Winter provide the ability to tie a text analyzing the underlying music or the score itself to specific sections of music. [CD-3] [CD-4] [CD-9]. Some of the most interesting and influential works were developed from scratch for CD-ROM. These include interactive "games" such as Laurie Anderson's "Puppet Motel" [CD-40][2] and the Resident's "Freak Show"[CD-30][3]. While some of the Voyager CD-ROMs have aged well, others, such as those that utilize video, have not. Most suffer from the limitations of the underlying programming technology which makes them appear relatively primitive by today's standards. Nevertheless, the Voyager CD-ROMs represent an important and pioneering body of work.

The first Voyager CD-ROM, and possibly the first consumer CD-ROM, was an interactive version of Beethoven's Ninth Symphony designed by Robert Winter, a UCLA music professor, and originally released in 1989. (An on-line demo is available [32].) As with other Voyager interactive music CD-ROMs, this was a hybrid disk consisting of a single data track and multiple audio tracks. The interactive component was a HyperCard stack. HyperCard was an immensely influential product of Apple that allowed non-programmers to create interactive programs [29][4].

Other hybrid CD-ROMs include "I Photograph to Remember" (IPTR) [CD-2] and "All My Hummingbirds Have Alibis" (ALLMY) [CD-5]. Pedro Meyer, the author of IPTR claims that this was the first CD-ROM with continuous sound and music ever produced. [19]. Although the work is

[2]An on-line demo is available [1]

[3]Lynn Ginsburg, writing in WIRED claimed Freak Show was "hailed as the best CD-ROM ever" [10]

[4]HyperCard was also used to create the immensely popular adventure game Myst (by the Cyan company) which also works well with the technologies described in this paper.

now available on the web, the CD-ROM provides an interesting opportunity to study the limits of CD-ROM technology. ALLMY is an experimental composition of three "imaginary ballets" composed by electronic music pioneer Morton Subotnick based upon illustrated novels by the surrealist Max Ernst [6].

Another historically significant Voyager publication is "Who Built America" [5], which supplemented the original printed text with multi-media materials [CD-16], because of a censorship attempt allegedly made by Apple Computer because of its treatment of homosexuality, birth control, and abortion [22].

Space restrictions preclude a more in-depth discussion of the Voyager Company materials – we trust this brief discussion has helped to illustrate why preservation of such materials is valuable. As we show through the remainder of this paper, preservation of a large collection of CD-ROMs does not require large-scale technological development.

4. EMULATION

The primary tool used for the research discussed in this paper was SheepShaver, an open-source and multi-platform Macintosh emulator. SheepShaver models the later PowerPC based classic Macintoshes executing System 7.5.2 through 9.0.4. A related program, BasiliskII models the earlier 600xx based Macs running all operating systems up to 8.1. In this section, we describe these emulation platforms and extensions we made to support mixed-mode CD-ROMs as well as network support for shared emulation environments.

Since the Voyager CD-ROMs were primarily published 1991-1997, we expected that any of the 7.x operating systems would be capable of supporting the Voyager CD-ROMs and previous work by Martin [16] had provided clear indication that later operating systems are probably incompatible with some of the Voyager CD-ROMs. In testing the Voyager CD-ROMs, we initially used Basilisk II executing System 7.6 (the final, and hence most refined version of the 7.x series); however, we encountered some performance issues (notably executing "A Hard Days Night" [CD-12]) and switched to SheepShaver, which has some significant performance enhancements, executing System 7.6.[5] With the addition of code to support mixed-mode CD-ROMs, all of the Voyager CD-ROMs tested appear to execute correctly in SheepShaver.[6] Testing consisted of installing and executing each CD-ROM in our emulation environment. We did not attempt to completely execute every aspect of every CD-ROM, but rather to determine with reasonable confidence that they behaved correctly.

Many of the Voyager CD-ROMs (and CD-ROMs in general) require some installation before they can be used. We believe this installation process represents a serious barrier to casual users. Hence we have investigated ways to create custom configurations for each unique CD-ROM that an end-user could download and execute. An emulator configuration consists of three components: a binary image of one of the Mac system ROMs, a disk drive image containing an installation of a Mac operating system, and a "preference" file defining the configuration of the emulated machine. The preference file defines the locations of the ROM image, the disk drive image, and any CD-ROM images, as well as other hardware parameters. For each of the Voyager CD-ROMs we created a separate preference file. After starting the emulator with a generic installation of System 7.6, we performed any necessary CD-ROM installation. The resulting hard disk image and preference file together form a configuration that could be cloned for end-users. When coupled with a distributed file-system supporting a global name space, there is no need to copy CD-ROM images to an end-users machines. For example, a preference file linking to one of the Voyager CD-ROMs (e.g. "All My Hummingbirds...") would contain an entry connecting a virtual CD-ROM drive to the corresponding image (`allmy.cue`).

```
cdrom /afs/iu.edu/home/\
       projects/sudoc/Voyager/allmy.cue
```

Cloning the hard drive image requires copying it to the end-user machine (roughly 75Mbytes), or, using a technology we describe later, copying a small "shadow copy" (a few Kbytes) that records only the changes to a networked reference copy. Before describing that technology, it's important to understand the architecture of these emulators and the steps required to extend them.

SheepShaver and BasiliskII are examples of "para-virtualized" emulators[7] which depend upon modifications to the underlying operating system for their operation. The most complicated aspect of system emulation is accurately modeling the I/O hardware such as the video monitor, network interface, and storage devices. Para-virtualization provides a significant simplification by modifying those portions of system software that access I/O hardware to allow that functionality to be provided by the host system without the need to model the target system hardware. In the case of SheepShaver and BasiliskII, para-virtualization is achieved by "patching" the ROMs that provide low-level hardware access.

SheepShaver and BasiliskII are organized approximately as illustrated in Figure 1. A processor model executes native Mac instructions from the application, the operating system, or the patched ROM. Any system call that accesses a device supported by the ROM is diverted to host code implementing generic versions of these devices. For example, SheepShaver provides a single CD-ROM driver. Any of the system calls specified in Apple "Technical Note DV22" [2], which defines the CD-ROM driver calls, is serviced by

[5]System 7.5.5 is available for free download from Apple and hence would be a good choice for implementing a virtual library using the techniques discussed in this paper.

[6]A possible exception is Blam! [CD-10], which was designed to both take charge of the user's computer and to be annoying. It is hard to be sure if the behavior of this title under emulation is annoying by intent or by virtue of imperfect emulation. Furthermore, a few of the experimental collections ([CD-37],[CD-45]) can best be described as "fragile".

[7]The distinction between emulation and virtualization can be quite fuzzy. Even systems supporting "true" virtualization generally must emulate I/O devices and often partially emulate the processor itself.

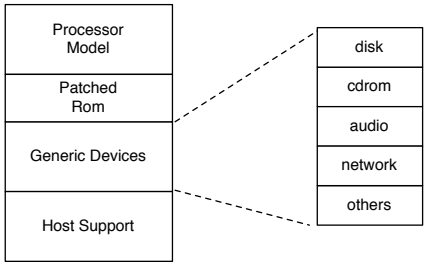


Figure 1: SheepShaver and BasiliskII Architecture

SheepShaver's generic CD-ROM driver. This, in turn, is supported by host specific code which may access host devices (for example the host CD-ROM hardware) or simulate the equivalent hardware.

Images of hybrid CD-ROMs (indeed audio CD-ROMs) are not supported by most emulation or virtualization environments because the audio track are generally handled directly by a CD-ROM which plays audio under the control of the host processor. In order to support images of hybrid CD-ROMs (discussed in Section 5) we created code that provides a simulated CD-ROM drive for these images. This simulation includes access to the host audio device. Modifications were required to both the emulator subsystem as well as to the virtual CD drive. Although this code is now part of the standard distribution for SheepShaver and BasiliskII, hybrid CD-ROM images are currently supported only on Unix (Linux) and OS X. Adding support for Windows hosts will be considerably easier and can directly utilize the primary module we created.

As mentioned, both emulators simulate the processor (CPU). The simplest, although slowest way to do this is through an interpreter that decodes and evaluates each target machine instruction as it is executed. This approach tends to suffer from poor performance. Both emulators accelerate processor simulation by on-the-fly translation of target machine code to host machine code (a so-call JIT). Thus, performance of both platforms is good on most code. SheepShaver also provides acceleration for QuickDraw (the classic Mac's graphics code) which we believe is the reason it did not have problems with "A Hard Day's Night."

SheepShaver and BasiliskII both use files containing binary images to represent system hard-drives. These hard-drive images need to be customized for many of the CD-ROMs to ensure the end-user does not have to navigate software installation procedures. Further, the hard-drive images should be cloned for each user so they modify only a local copy. The use of such hard-drive images imposes a significant start-up time in copying them to the end-user's workstation. To eliminate this source of delay, we integrated libvhd from the Xen [34] distribution with SheepShaver and BasiliskII. This modification enables the use of hard-drive images in the Virtual Hard Disk (VHD) format used by Microsoft Virtual PC, Xen, and VirtualBox. [20, 34] VHD images can be layered with a base image and layers providing "differences." Using such layered hard-drive images, the base image may remain on the distributed file system while the difference image is copied to the end-user machine. This difference image is initially only a few KBytes. As the end-user uses a particular image, any disk writes are made only to the local difference image and the base image remains unchanged. Thus, copying a usable emulator configuration to a patron's workstation involves moving only a few Kbytes consisting of a configuration file and a virtual hard drive difference image from the collection server. Both the CD-ROM image and base virtual hard drive remain on the server and any required data is copied "on-the-fly" as the patron executes the emulator.

In general, we found SheepShaver and BasiliskII to be moderately stable platforms; however, both suffer from unexpected crashes. A close examination of the code base betrays its history as a open-source project which has resulted in a sprawling code base supporting many host platforms in a large variety of configurations. However, the basic architecture is sound and, with a modest investment, they could become viable platforms for preservation of Mac CD-ROMs.

5. CD-ROM IMAGES

Most CD-ROMs can be saved as bit-faithful images that can be used in place of the physical media – an important exception, which does not apply to the Voyager CD-ROMs, are copy protected CD-ROMs implemented with violations to CD standards. A CD-ROM image consists of a binary file containing the "user data" of the original CD-ROM along with a "table of contents" file replicating the critical metadata from the CD-ROM. In an emulation environment, with an appropriate "device model", this image can be used in a manner that is indistinguishable from the original CD-ROM.

The Voyager CD-ROMs are based upon the Audio CD standard (the "Red Book") extended by the "Yellow Book."[8] CD-ROMs are recorded as a sequence of fixed sized (2352 byte) sectors which may contain either audio data or user data – in the former case, a sector contains 1/75 second digitally encoded sound; in the latter case a sector contains 2048 bytes of data protected by additional error correcting bits. CD-ROMs encode metadata in parallel with the sector data through a mechanism called "subchannels." These metadata include information about the organization of the disk into tracks – equivalent to a song in an audio CD – as well as basic cataloging information about the disk. By convention, sectors are numbered according to their temporal position from the beginning of the CD-ROM in units of minutes, seconds, and 1/75 second – even in pure data CD-ROMs where time has no real meaning.

The Voyager CD-ROMs fall into two (physical) categories. Most consist of a single "track" containing an Apple (HFS) file system and in practice behave as read-only disk drives. A small number are mixed-mode CD-ROMs (later called Enhanced CD) with multiple tracks – a data track containing an HFS file system followed by one or more audio tracks. The data-only CD-ROMs present no problems for SheepShaver or BasiliskII in either their physical or image forms. The mixed-mode CD-ROMs do not work on most

[8]A good overview is provided by the ECMA-130 standard that parallels the Yellow Book [8].

platforms in their physical form and, prior to our work, were not supported as images.

Imaging either type of CD-ROM is fully supported by existing open-source tools. In a Linux or OS X environment, data-only CD-ROMs can be imaged by the Unix program `dd`. For example, the following command will create a binary image named "cd.bin" of a CD-ROM in the drive /dev/cdrom.

```
dd if=/dev/cdrom of=cd.bin
```

In SheepShaver or Basilisk, this file may be "mounted" by including the following in the preference file read at start up.

```
cdrom cd.bin
```

Similarly, most PC virtualization tools such as VMWare and VirtualBox support such binary files directly, although the operating systems running on these tools cannot generally read HFS file systems.

Existing open-source tools do support creating images of the Voyager mixed-mode CD-ROMs which can be executed in Sheepshaver using the extensions we developed. To create hybrid images we utilized the Linux tool `cdrdao`. For example, to create table of contents and data files `allmy.toc, allmy.bin` from a physical CD-ROM execute:

```
cdrdao read-cd --read-raw --device 1,0,0 \
       --datafile allmy.bin allmy.toc
```

It appears that none of the most widely used PC emulators (VMWare, Virtual Box, Xen, Qemu) support hybrid CD-ROMs. In addition to the Voyager materials, there were numerous hybrid audio discs published over a 10-year period.

6. A NOTE ON COPYRIGHT

Throughout this paper we have noted that copyright law may be the single greatest impediment to the preservation of born-digital materials. For the work we have described, there are three entities covered by copyright – the Macintosh ROM used by the emulator, the Macintosh Operating System, and the CD-ROMs themselves. Apple has made versions of the Mac operating system (7.5) publicly available, although the conditions of use are not clear. No version of the Macintosh ROM is freely available. At least these two system components have a single entity with which a library or archive would need to negotiate for creating a virtual CD-ROM collection. A similar situation exists for Windows based emulation, although the existence of commercial emulators suggests that, for the operating system and BIOS, clear use protocols exist. The problem with CD-ROMs is more complicated – the existing laws are unclear – yet it is unlikely that a library or archive could reasonably negotiate with the individual copyright holders. The following describes the laws in the United States as we understand them.

Under existing law 17 USC § 108, Libraries may make three copies of a work for preservation purposes provided they are not distributed in digital form outside the premises of the library or archives and this right is exercised to replace a work that is damaged, deteriorating, ..., or in an obsolete format. [13, 27]. In addition, it must be determined that an unused copy cannot be obtained at a fair price.

In practical terms, current copyright law limits the immediate application of the techniques described in this paper to libraries and archives who wish to provide their patrons with electronic access to materials which they own. In the long-term, our vision requires networked access to collections of such digital materials. The Section 108 Study Group [26] of the United States Copyright Office has recommended that:

> The prohibition on off-site lending of digital replacement copies should be modified so that if the library's or archives' original copy of a work is in a physical digital medium that can lawfully be lent off-site, then it may also lend for off-site use any replacement copy reproduced in the same or equivalent physical digital medium, with technological protection measures equivalent to those applied to the original (if any).

The Digital Millennium Copyright Act (DMCA) appears to complicate the situation where breaking copy protection is a necessary step in creating replacement copies; however, for libraries and archives a preservation exemption has been created for programs or video games distributed in obsolete formats requiring original media or hardware as a condition of access [4].

A comprehensive international survey of the copyright issues relating to digital preservation was created by the Library of Congress and others [15].

7. DISCUSSION

The major conclusion of this work is that CD-ROMs published for the Classic Macintosh can probably be preserved for access by future generations using existing technology and a networked archive could be created using the additional techniques described in this paper. The specific combination of emulator and operating system, SheepShaver executing System 7.6, proved to be a very good platform for accessing the Voyager Company CD-ROMs. We successfully tested 48 of the approximately 75 published Voyagers CD-ROMs including instances from the all the years of publication. The full set of tested CD-ROMs is provided in Appendix A. This table also provides a summary of the software provided on each CD-ROM, for example, HyperCard or MacroMind Director and hence some indication of the technologies required. Most required some form of installation ranging from copying fonts to the system folder to running a provided installer. None required externally provided software in System 7.6; although, when executed under later versions of the Mac operating system they may. Note that a handful of titles – most notably those that were compilations of experimental multi-media art – crash easily. In contrast, the more commercial titles are quite robust.

We believe that many of the Classic Macintosh CD-ROMs produced by other publishers can also be preserved using the techniques we have described. Although we did test a small number of other titles including the wildly popular game "Myst", our confidence comes from the relatively tight control Apple exerted over its platform during the Classic period.

The SheepShaver platform performs well and is generally easy to use; however, it can be more fragile than is desirable. For example, hard crashes are more common than we recall from our past use of classic Macintoshes. We believe this could be improved with some engineering effort. As mentioned previously, existing emulation tools can support most of the Voyager CD-ROMs with the notable exception of those utilizing mixed-mode data/audio CD formats; we have described emulator extensions to support these.

As we have shown, all of the pieces exist to enable a virtual collection of the Voyager Company CD-ROMs – from distributed file systems to enable sharing across institutions, to the required emulation tools, to any required software. We have successfully extended SheepShaver to support hybrid disk images and to support virtual hard drives which will allow end-users to down-load very small CD-ROM specific configurations. On the OS X environment, these configurations open through mouse clicks which makes their use especially simple; this functionality could be added to the Windows and Linux environments. The code we have added is now part of the standard SheepShaver source distribution; there remains a modest task of extending this to the Windows platform. This approach also represents a significant architectural alternative to the technique we previously used to deliver customized emulation environments. Our approach does ultimately depend upon preserving the underlying emulator; the KEEP project has as its aim the preservation of such environments. [14]

One of the anonymous referees asked why we don't use Remote Desktop or VNC to offload the client's workstation completely. The short answer is these technologies do not work well with multimedia. For example, VNC doesn't support audio and no remote access technology we have tested works well with video.

While the overall message of this paper can be viewed a positive – no super-human technical effort is required to preserve an important slice of computer history – the clock is ticking. The software required is disappearing rapidly and the set of people with sufficient technical knowledge to bring the existing emulators up to "production quality" is declining. Finally, there are significant legal challenges to building an archive of commercial software.

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.

8. REFERENCES

[1] L. Anderson. Voyager puppet motel demo. `http://blip.tv/file/462115/`. Accessed October 2010.

[2] Apple Computer. CD-ROM driver calls. Technical Report DV22, Apple Computer, 1999. `http://developer.apple.com/legacy/mac/library/#technotes/dv/dv_22.html`. Accessed November 2010.

[3] Basilisk. The official BasiliskII home page, 2010. `http://basilisk.cebix.net`. Accessed October 2010.

[4] Copyright Office, Library of Congress. Exemption to prohibition on circumvention of copyright protection schemes for access control technologies. *Federal Register*, 71(227):68472–68480, November 2006.

[5] A. Darien. CD-ROM review who built America. *The Journal of Multi Media History*, 1(1), 1998. `http://www.albany.edu/jmmh/vol1no1/wba_review1.html`. Acessed November 2010.

[6] S. Davison. All my hummingbirds have alibis (a review). *Notes, Second Series*, 53(2):530–533, December 1996. `http://www.jstor.org/stable/900152`. Accessed November 2010.

[7] J. P. Eakin. Reading comics: Art Spiegleman on CD-ROM. Provided by the author, March 2009.

[8] ECMA. Standard ECMA-130 data interchange on read-only 120mm optical data disks (CD-ROM), June 1996. `http://www.ecma-international.org/publications/standards/Ecma-130.htm`. Accessed October 2010.

[9] S. Gilheany. Preserving digital information forever and a call for emulators. In *Digital Libraries Asia 98: The Digital Era: Implications, Challenges, and Issues*, 1998.

[10] L. Ginsburg. Twin peaks meets SimCity. *WIRED*, September 1995. `http://www.wired.com/wired/archive/3.09/residents.html`. Accessed October 2010.

[11] 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.

[12] 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.

[13] P. B. Hirtle. Digital preservation and copyright, 2003. `http://fairuse.stanford.edu/commentary_and_analysis/2003_11_hi\rtle.html` e.html.

[14] Keeping Emulation Environments Portable. `http://www.keep-project.eu/expub2/index.php.\AccessedSeptember2011`.

[15] Library of Congress National Digital Information Infrastructure and Preservation Program, The Joint Information Systems Committee, The Open Access to Knowedge (OAK) Law Project, The SURFfoundation. International study on the impact of copyright law on digital preservation, July 2008.

[16] J. Martin. Voyager company CD-ROMs: Production history and preservation challenges of commercial interactive media, 20?? `http://www.eai.org/resourceguide/preservation/`

computer/pdf-docs/voyager%_casestudy.pdf. Accessed October 2010.

[17] A. T. McCray and M. E. Gallagher. Principles for digital library development. *Commununications of the ACM*, 44(5):48–54, 2001.

[18] P. Mellor. CaMiLEON: emulation and BBC doomsday. *RLG DigiNews*, 7(2), 2003.

[19] P. Meyer. "...some background thoughts", May 2001. http://www.zonezero.com/exposiciones/fotografos/fotografio/work.html. Accessed October 2010.

[20] Microsoft. Virtual hard disk image format specification, October 2006. http://technet.microsoft.com/en-us/virtualserver/bb676673.aspx. Accessed November 2010.

[21] OpenAFS. OpenAFS, 2010. http://www.openafs.org. Accessed November 2010.

[22] J. F. Reynolds. Who built america controversy, March 1995. http://h-net.msu.edu/cgi-bin/logbrowse.pl?trx=vx&list=H-Mmedia&month=95%03&week=d&msg=bvNSxOEHMLiyFRbjGUu9Lw&user=&pw=. Accessed March 2011.

[23] J. Rothenberg. Ensuring the longevity of digital information. *Scientific American*, 272(1):42–47, January 1995.

[24] J. Rothenberg. An experiment in using emulation to preserve digital publications. Technical report, Koninklijke Bibliotheek, July 2000.

[25] J. Rothenberg. Using emulation to preserve digital documents. Technical report, Koninklijke Bibliotheek, July 2000.

[26] Section 108 Study Group. The Section 108 Study Group Report. Technical report, March 2008.

[27] Section 108 Study Group. About section 108, 2009. http://www.section108.gov/about.html. Accessed November 2009.

[28] Sheepshaver. The official SheepShaver home page, 2010. http://sheepshaver.cebix.net. Accessed October 2010.

[29] Smackerel. When multimedia was black and white, 2005. http://www.smackerel.net/black_white.html. Accessed November 2010.

[30] Variable Media. Seeing double emulation theory and practice, 2004. http:/www.variablemedia.net/e/seeingdouble/home.html. Accessed November, 2006.

[31] A. Virshup. The teachings of Bob Stein. *WIRED*, July 1996. http://www.wired.com/wired/archive/4.07/stein_pr.html. Accessed October 2010.

[32] R. Winter. The cd-companion to Beethoven's ninth symphony, 2009. Flash demo http://www.futureofthebook.org/blog/archives/2009/11/published_by_the_v%oyager_compa.html. Accessed October 2010.

[33] K. Woods and G. Brown. Assisted emulation for legacy executables. *The International Journal of Digital Curation*, 5(1):160–171, 2010.

[34] Xen. Xen hypervisor, 2010. http://www.xen.org/products/xen_source.html. Accessed November 2010.

[35] D. Young. Inventing interactive: Early hypercard creativity, April 2010. http://www.inventinginteractive.com/2010/04/15/early-hypercard-creativi%ty/. Accessed October 2010.

[36] D. Young. Inventing interactive: Voyager (1989-2000), Feburary 2010. http://www.inventinginteractive.com/2010/02/22/voyager-1989-2000/. Accessed October 2010.

A. TEST RESULTS

Title	Year	H	P	F	D	M
1. Baseball's Greatest Hits	1991	✓		✓		
2. I Photograph to Remember	1991				✓	✓
3. Ludwig van Beethoven Symphony No. 9 (multilanguage edition)	1991	✓		✓		✓
4. Mozart String Quartet in C Major	1991	✓		✓		✓
5. All My Hummingbirds Have Alibis	1992			✓	✓	✓
6. Bach and Before (So I've Heard Vol. 1)	1992	✓		✓		✓
7. Classical Ideal (So I've Heard Vol. 2)	1992	✓		✓		✓
8. Poetry in Motion	1992	✓				
9. Richard Straus Three Tone Poems	1992	✓		✓		✓
10. Blam! [9]	1993	✓			✓	
11. Children's Songbook	1993				✓	
12. A Hard Day's Night [10]	1993	✓		✓		
13. Hikaruhana (Shining Flower)	1993				✓	
14. Planetary Taxi	1993	✓		✓		
15. Take Five [11]	1993			✓	✓	
16. Who Built America	1993	✓		✓		
17. American Poetry	1994	✓				
18. Antonín Dvořák Symphony No. 9	1994	✓		✓		✓
19. Beethoven and Beyond (So I've Heard Vol. 3)	1994	✓		✓		✓
20. Comic Book Confidential [12]	1994	✓		✓		
21. Criterion Goes To the Movies	1994	✓		✓		
22. Dazzeloids	1994		✓			
23. Defending Human Attributes in the Age of Machines	1994	✓		✓		
24. Ephemeral Films	1994		✓			
25. Exotic Japan	1994				✓	
26. The First Emperor of China	1994		✓			
27. For All Mankind	1994		✓			
28. Macbeth	1994	✓				
29. Maus a Survivors Tale	1994	✓		✓		
30. The Residents Freakshow	1994				✓	
31. The Society of Mind	1994	✓		✓		
32. Stephen Jay Gould on Evolution	1994	✓		✓		
33. Amnesty Interactive	1995		✓			
34. Day After Trinity	1995		✓			
35. Live From Death Row	1995		✓			
36. Morton Subotnick's Making Music [13]	1995		✓			
37. New Voices New Visions [14]	1995	✓			✓	
38. Our Secret Century (Volumes 1-4) [15]	1995		✓			
39. Poetry in Motion II	1995		✓			
40. Puppet Motel	1995				✓	
41. Starry Night	1995				✓	
42. Theatre of the Imagination	1995				✓	
43. Truths & Fictions: A Journey from Documentary to Digital Photography	1995				✓	
44. With Open Eyes	1995				✓	
45. New Voices New Visions 1995 [16]	1996	✓			✓	
46. Sacred and Secular: The Aerial Photography of Marilyn Bridges	1996				✓	
47. Witness to the Future	1996			✓		
48. Fun With Architecture	1997			✓		

Key: **H** uses Hypercard, **P** includes custom program, **F** includes custom fonts, **D** MacroMind Director, **M** Mixed-mode CD-ROM.

[9] "Hangs" for long periods. May be intended behavior.
[10] Doesn't perform adequately in BasiliskII.
[11] Seems a bit fragile – doesn't restart properly without reboot.
[12] Movie crashes unless version of hypercard on CD-ROM is used.
[13] MIDI interface is not supported in emulator, but is not required.
[14] Collection of multimedia art from competition – some of these crash easily.
[15] Uses Oracle media objects.
[16] Second collection of multimedia art from competition – some of these crash easily.

A Braille Conversion Service Using GPU and Human Interaction by Computer Vision *

Roman Graf
Digital Memory Engineering
Safety & Security Department
AIT Austrian Institute of Technology GmbH
Vienna, Austria
roman.graf@ait.ac.at

Reinhold Huber-Mörk
High Performance Image Processing
Safety & Security Department
AIT Austrian Institute of Technology GmbH
Seibersdorf, Austria
reinhold.huber@ait.ac.at

ABSTRACT

Scalable systems and services for preserving digital content became important technologies with increasing volumes of digitized data. This paper presents a new Braille converter service that is a sample implementation of scalable service for preserving digital content. The converter service facilitates complex conversion problems regarding Braille code. Braille code is a method which allows visually impaired people to read and write tactile text. Using a GPU with the CUDA architecture allows the creation of a parallel processing service with enhanced scalability. The Braille converter is a web service that provides automatic conversion from the older BRF to the newer PEF Braille format. This service can manage a large number of objects. Speedups on the order of magnitude of 5000 to 6900 (depending on the size of the object) were achieved using a GPU (GTX460 graphics card) with respect to a CPU implementation. An extension involving an image processing system is used for human interaction. Optical pattern recognition allows Braille code creation using Braille patterns. No special input device and skills are needed, only familiarity with Braille code is required.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: System issues; I.5.5 [**Pattern Recognition**]: Interactive systems; H.3.5 [**Online Information Services**]: Web-based services

General Terms

Algorithms

*This work was supported in part by the EU FP7 Project SCAPE (GA#270137) `www.scape-project.eu`. We would like to thank Susan Jolly from `www.dotlessbraille.org` for providing useful information regarding Braille formats.

Keywords

digital preservation, CUDA, performance measurement, information retrieval, image processing, haptic I/O

1. INTRODUCTION

We describe a scalable Braille conversion web service using CUDA GPU parallel processing. CUDA is a technology developed by NVIDIA [9] and could be used by a conversion service to increase performance and enhance scalability. CUDA programming requires a specific hardware. In this work, a GTX460 graphics card was used.

Scalability plays an important role with increasing volumes of digitized data. One of the goals of the SCAPE project is to create scalable services for digital preservation of large amounts of data. The SCAPE platform will include a storage infrastructure and an execution environment for performing sustainable, data-intensive and scalable digital preservation activities through parallel processing.

The Braille code [4] conversion service available through a web server is a sample implementation for such a scalable service. There is a lack of public Braille services and still no proper solution to the problem of automating Braille conversions [6]. Automatic migration of Braille files in BRF file format [3] to the newer and more flexible PEF format is a practical implementation of a service that needs scalability. The output of the conversion service is a PEF file which can further be used for embossing or for presentation using standard Braille display. A demo workflow was implemented based on the Planets Workflow Engine [10]. Customized workflows support the definition of service parameters that manage workflow execution.

Braille [4] is a system which enables visually impaired persons to read text from tactile patterns by touch. The same system is also suitable for writing tactile text. The Braille code is a set of tactile patterns combined from raised dots. A physically tactile pattern is presented as a cell of six dots arranged in two columns and three rows. Letters, numerals and punctuation can be represented using different dots combinations.

Interactive and automatic Braille recognition from images was successfully performed with different acquisition setups, various algorithms and different output format and media [2], [8]. We will describe a system with a fixed camera, edge based segmentation and recognition of Braille characters.

This paper is organized as follows. Section 2 describes the challenges associated with Braille encodings. Section 3 introduces a web-service for conversion of Braille between

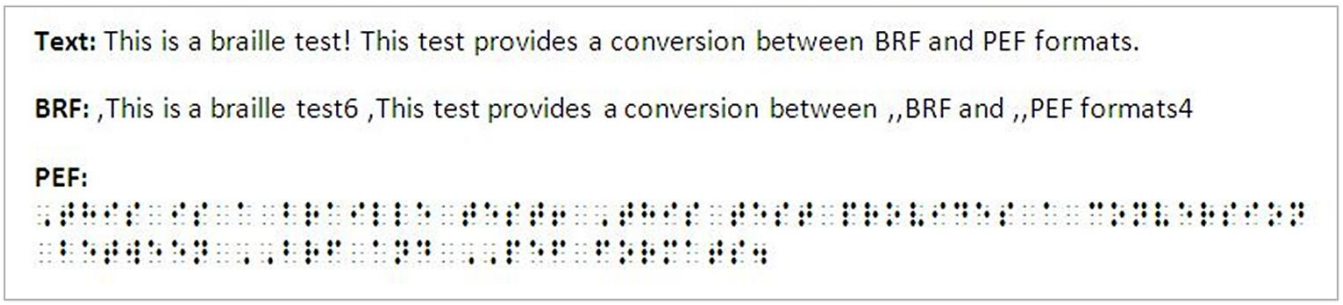


Figure 1: Representation of a sample text in BRF and PEF format.

BRF and PEF encodings and Sec. 4 demonstrates scalability based on GPU processing. Section 5 describes image based Braille pattern recognition. Section 6 concludes the paper.

2. BRAILLE CODING CHALLENGES

There are a number of different formats for digital representation of Braille, two of which are considered in this section. A BRF file [3] is an ASCII file where the ASCII characters simply transliterate the Braille cells according to some convention (consider the the North American ASCII Braille BRF example in Fig. 1).

The more recent format for Braille files called PEF (Portable Embosser Format) [5] was developed in 2005 by the Swedish Braille library. PEF is not-yet widely adopted but has some advantages when compared to the BRF file format because it contains information about file content, proper Braille publishing standard, file sharing ability, and long term archive preservation safety. PEF files have a header which references the print source and other important metadata like title, author and so on. Customized metadata are also possible. The PEF is a document type that represents Braille pages in digital form, accurately and unambiguously; regardless of language, location, embosser settings, Braille code and computer environment. PEF uses Unicode Braille patterns which are widely accepted as a part of Unicode standard.

Braille files are Braille translations [3] of printed texts produced manually by experienced translators. Braille coding utilizes various combinations of contractions, markup, direct representation, and whitespace formatting. Each language has one or more different Braille codes for converting text to Braille (literature, technical material, music, computer and so on). Currently there is no way to uniquely identify which Braille system has been used to produce the Braille file. Furthermore, a Braille code is characterized by a context-sensitive grammar and, even if we know the correct specification for the used Braille system, it is impossible to regenerate the print text completely accurate. ASCII Braille represents Braille cells by ASCII characters instead of Unicode and has an advantage that it is easier for humans to use. The disadvantage of ASCII Braille is that the encoding has to be defined. The Unicode advantage is that it is an international standard.

3. WEB SERVICE FOR CONVERSION BETWEEN BRF AND PEF FORMATS

The workflow engine provides the functionality of Braille data conversion from the BRF format to the PEF format using a predefined conversion workflow and the Planets digital object model [10]. The functionality to manage Planets digital objects is provided through the Braille conversion web service written in Java and deployed with the JBoss application server. This service implements methods that allow the workflow engine to read data from a BRF file, to convert it and to write data to a PEF file.

The Braille conversion use-case describes the performed activities during the processing of the normalization strategy. The main goal of this service is the conversion of the source content from its original BRF data format into an open, preservation-friendly and compatible, PEF format. The conversion service performs the following actions:

1. The use-case starts with a user call of the conversion service providing the Braille data of a collection in BRF.
2. The service generates a preservation plan for each item in the data collection.
3. The normalization strategy processing starts with BRF content evaluation. Binary files of the processed item will be harvested based on their URL. Information is collected from content providers and integrated into a representation of the objects in the preservation tasks.
4. Metadata is evaluated in order to build a PEF file header. Expected header should contain following terms: ”title”, ”date”, ”format”, ”description”, etc.
5. Create error report.
6. Run the migration accordingly to the preservation plan.
7. Store migration results into the PEF file.
8. Generate report.

4. SCALABILITY ENHANCEMENT USING GPU

Scalability could be achieved by parallel processing, e.g. using OpenMP on multicore processors, distributed processing or using OpenCL or CUDA on general purpose graphics cards. In this work we describe a CUDA [9] implementation which utilizes GPU parallel processing. The implementation extends the Braille conversion service in order to enhance scalability. In an experimental setup BRF files of different sizes were converted to PEF files using a GPU processing application. In the experiment two implementations of the conversion service are compared. One implementation uses traditional CPU processing whereas the second implementation uses GPU parallel processing. Pure calculation processing time and total processing time were measured.

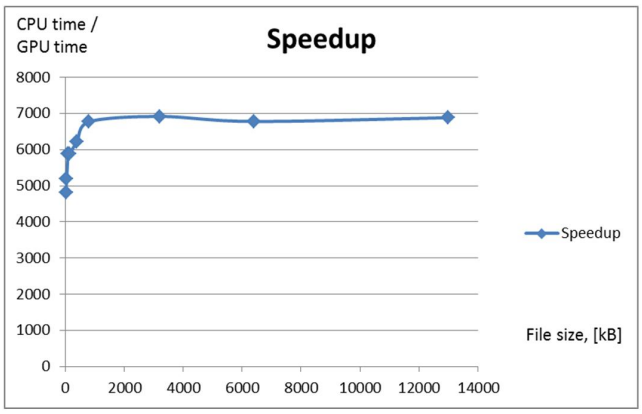


Figure 2: Speedup achieved using GPU parallel processing.

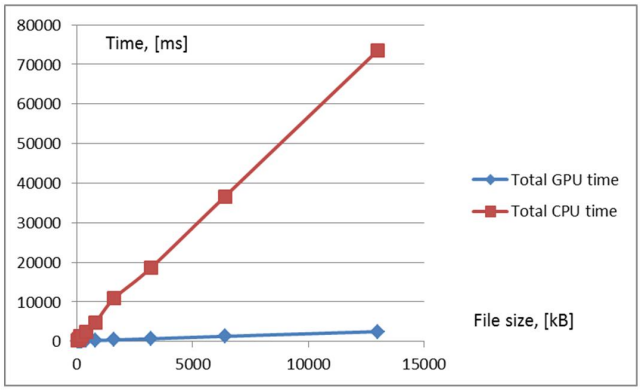


Figure 3: Dependency between file size and execution time.

In order to prove the advantages of GPU parallel processing the input data from the BRF file was divided into chunks and loaded to the device memory. In order to map algorithms to the GPU each chunk represents a BRF element that can be converted to PEF Unicode values using a conversion table. Then chunks are processed in parallel by the GPU kernels and results copied back to the host. Parallelization, i.e. thread and memory management, is provided by CUDA. Each natural language requires a special conversion algorithm. In the experiment a conversion algorithm for American English was implemented. The resulting output data is written to a PEF file. Performance measurements were computed to evaluate GPU and CPU processing times (used graphics card: GTX460). The relation of the pure GPU processing time in respect to pure CPU processing time, see Fig. 2, reveals the application processing time without taking in account memory management time. Figure 3 indicates the relation of the total processing times for both implementations including memory management from the start of the conversion service calculations to the completion of the content conversion process.

Files of larger sizes achieve higher speedups, where file size is measured in KB and time in milliseconds. Figure 3 depicts the dependency between the time needed for the BRF file content reading, converting and writing to PEF file using GPU and the time consumed by the same operations using

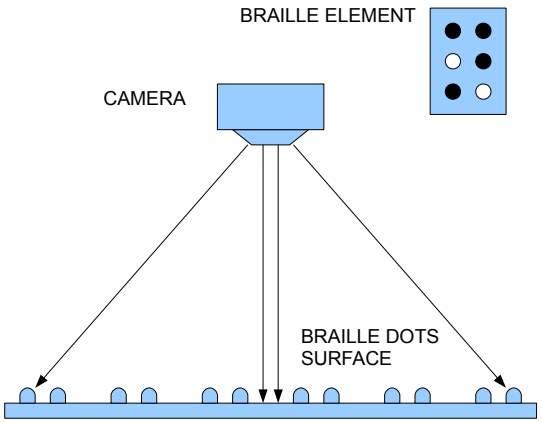


Figure 4: Image acquisition of Braille elements.

CPU. The memory management and file read/write operations have been taken into account. In order to evaluate the dependency of performance on file size a sample of eleven files having sizes between 25KB and 13MB were converted. The migration process on CPU mostly takes more time as the content conversion process on GPU. With the larger file size the CPU time consumption increases almost proportionally but the GPU time consumption remains approximately the same.

The gain from using the GPU starts with a file size of about 50KB because for smaller files, the overheads associated with allocating GPU memory dominate the computation time. While GPU parallel processing achieves its maximum speedup (6900 times) for Braille calculations for the file sizes starting from about 13MB. The parallel implementation can be improved by overlapping communication with computation. The reason for the significant speedup is that calculation task normally (with CPU) computed sequentially was broken down to the sub tasks and processed in parallel. The GPU comprises hundreds of cores (336 for GTX460 graphics card). The higher is the cores number the higher is the achieved speedup. Each core supports multiple threads that can be executed in parallel. Each thread has a subtask to execute. Therefore more resources are devoted to data processing rather than data caching and flow control. Web service and workflow engine overhead time consumption is independent from conversion time.

5. EXTRACTION OF BRAILLE CODE FROM IMAGES AND CONVERSION

This section describes an extension to the presented conversion service. Image processing is applied to Braille code extraction. Images are acquired using an area camera mounted at some fixed distance to the surface holding the Braille elements as shown in Fig. 4. Perspective distortions are avoided by the setup and geometric lense distortion is corrected. The camera provides gray-scale images of the surface holding physical Braille patterns.

The goal of this extension is to enable Braille coding for the users which are not familiar with Braille devices currently used as a computer interface. Currently, Braille code creation in digital form is only possible using specific Braille input devices. By interaction with a computer vision setup a user could arrange Braille pattern manually using physical

Braille elements, i.e. physical building blocks, representing Braille codes. The user identifies the meaning of each element scanning tactile pattern by moving the finger upon it. Visual impaired people are familiar with this technique and use it for Braille reading of printed books. Once Braille code creation is completed on a predefined surface the user starts imaging of the surface holding the Braille patterns. Subsequently, using a pattern recognition algorithm pins and semantics of Braille patterns are identified and delivered in ASCII code or Unicode standard. The output of pattern recognition and conversion provides an input for Braille conversion service described in Sec. 3.

In order to extract Braille elements in the image Braille and raster dots are initially segmented from the background [11]. Point and edge based features are regarded to be more robust against lighting variants [7]. In the suggested method the Canny edge detection algorithm was employed [1] in the segmentation step. The Braille pattern dots (see Fig. 1) used in the experiment consists of black points on a white surface. In the experimental setup the dot diameter is about 10 pixels. The placeholders for empty Braille pattern dots are depicted as smaller circles. The placeholder diameter is about 6 pixels. The placeholder dots are also important in Braille pattern calculation. The separation between Braille and placeholder dots is based on expected Braille dot height, width and maximal pixel count.

The Braille pattern recognition and conversion algorithm is summarized as follows:

1. Retrieve a Braille pattern image and apply geometric undistortion.
2. Segmentation of the retrieved image using the Canny edge detector. The output of segmentation is an array of detected points.
3. Verification of detected placeholder points using a Braille code grid, as placeholder dots are more accurately localized than Braille dots. A suitable maximal pixel count value for dot discrimination using the described setup was found to be 14.
4. Remove false positive detections. Extracted detections that do not match the predefined grid are removed.
5. Merge spatially adjacent placeholder dots from step 3. This step rejects pixels that are part of already detected Braille dots.
6. Detect Braille dots using the grid derived from placeholder dots.
7. Compute Braille patterns from detected Braille points. Braille patterns (for example 1-2-5) that are suitable for further processing are obtained.
8. Compute ASCII Braille code or Unicode values.

The resulting Braille encoding is used as input to the web conversion service described in Sec. 3.

6. CONCLUSION

A new scalable open Braille conversion web service for preserving digital content was created. The service provides conversion of Braille files into the new PEF format from the widely spread BRF format. Braille conversion service scalability is improved by application of CUDA GPU parallel processing. Measurements of conversion times for different BRF file sizes and comparison of results for GPU and CPU processing were given. Speedup enhancements up to 6900 times were achieved. The GPU parallel processing efficiency depends on the file size. For example 50KB file achieves speedup about 5000 times whereas 13MB file speedup is more than 6900 times. Apparently GPU processing is more efficient then CPU processing in terms of Braille conversion for large file sizes (50KB - 13MB) and enhances scalability of conversion service for large files collections.

The scalability improvement is that BRF files in the range between 50KB and 13Mb can be more efficient converted to the PEF files athrough GPU parallel processing and a web service. This acceleration is needed to efficiently migrate large amounts of currently widely used BRF file archives into PEF. Future work will include evaluation of scalability involving larger collections.

Image processing and pattern recognition can be used to enable Braille coding without Braille input device and to create input data for the Braille conversion service. Images of physical Braille patterns are acquired and automatically recognized and can also make use of the described web conversion service.

7. REFERENCES

[1] Canny, J.: A computational approach to edge detection. IEEE Trans. Pat. Anal. Mach. Intell. 8(6), 679–698 (1986)

[2] François, G., Calders, P.: The reproduction of Braille originals by means of optical pattern recognition. In: Proc. Int. Workshop on Computer Braille Production. pp. 119–122 (1985)

[3] Frees, B., Strobbe, C., Engelen, J.: Generating braille from Openoffice.org. In: Proc. Intl. Conf. Computers helping people with special needs. LNCS, vol. 6179, pp. 81–88 (2010)

[4] Jiménez, J., Olea, J., Torres, J., Alonso, I., Harder, D., Fischer, K.: Biography of Louis Braille and invention of the Braille alphabet. Survey of Ophthalmology 54(1), 142–149 (2009)

[5] Leas, D., Persoon, E., Soiffer, N., Zacherle, M.: Daisy 3: A standard for accessible multimedia books. IEEE Multimedia 15(4), 28–37 (2008)

[6] Manohar, P., Parthasarathy, A.: An innovative Braille system keyboard for the visually impaired. In: Proc. of UKSim: Intl. Conf. on Comp. Modelling and Simulation. pp. 559–562 (2009)

[7] Marr, D., Hildreth, E.: Theory of edge detection. Proc. of the Royal Soc. London B-207, 187–217 (1980)

[8] Mihara, Y., Sugimoto, A., Shibayama, E., Takahashi, S.: An interactive Braille-recognition system for the visually impaired based on a portable camera. In: Proc. of CHI'05 extended abstracts on Human factors in comp. systems. pp. 1653–1656 (2005)

[9] Owens, J.D., Luebke, D., Govindaraju, N., Harris, M., Krüger, J., Lefohn, A.E., Purcell, T.: A survey of general-purpose computation on graphics hardware. Comput. Graph. Forum 26(1), 80–113 (2007)

[10] Schmidt, R., King, R., Jackson, A.N., Wilson, C., Steeg, F., Melms, P.: A framework for distributed preservation workflows. Intl. J. of Digital Curation 5(1), 205–217 (2010)

[11] Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. J. Electron. Imaging 13(1), 146–165 (2004)

Evolving Domains, Problems and Solutions for Long Term Digital Preservation

Orit Edelstein
IBM Research - Haifa
Haifa, Israel
edelstein@il.ibm.com

Michael Factor
IBM Research - Haifa
Haifa, Israel
factor@il.ibm.com

Ross King
AIT Austrian Institute of Technology GmbH
ross.king@ait.ac.at

Thomas Risse
L3S Research Center
Hannover, Germany
risse@L3S.de

Eliot Salant
IBM Research - Haifa
Haifa, Israel
salant@il.ibm.com

Philip Taylor
SAP (UK) Ltd.
SAP Research
Belfast, Northern Ireland
philip.taylor@sap.com

ABSTRACT

We present, compare and contrast new directions in long term digital preservation as covered by the four large European Community funded research projects that started in 2011. The new projects widen the domain of digital preservation from the traditional purview of memory institutions preserving documents to include scenarios such as health-care, data with direct commercial value, and web-based data. Some of these projects consider not only how to preserve the programs needed to interpret the data but also how to manage and preserve the related workflows. Considerations such as risk analysis and cost estimation are built into some of them, and more than one of these efforts is examining the use of cloud-based technologies. All projects look into programmatic solutions, while emphasizing different aspects such as data collection, scalability, reconfigurability, and full lifecycle management. These new directions will make digital preservation applicable to a wider domain of users and will give better tools to assist in the process.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software

Keywords

Preservation, Web Archives, Software as a Service, Business Processes

1. INTRODUCTION

This paper presents the directions of the newly started major efforts on long term digital preservation partially funded by the European Union's FP7 initiative. There are four largest projects (overall budget of above eight million Euro each) funded by the EC on long term digital preservation that started in the last year.

While all four project address digital preservation, they differ in what data are being preserved, how the data are identified, and how the data are preserved. All of these projects consider and, when appropriate, use results of previous digital preservation projects.

We discuss the motivation and objectives of the four efforts, the target communities, and the respective stakeholders. The solutions chosen are presented and alternatives are discussed. By comparing the four projects, highlighting the areas where they complement each other, where they contrast, and what they cover, we are reporting the extent of the current effort within FP7 and their expected contribution to the domain of long term digital preservation.

The four projects are:

- ARCOMEM[1] - From Collect-All Archives to Community Memories - is about memory institutions like archives, museums and libraries in the age of the social web. Social media are becoming more and more pervasive in all areas of life. ARCOMEM's aim is to help to transform archives into collective memories that are more tightly integrated with their community of users and to exploit Web 2.0 and the wisdom of crowds to make web archiving a more selective and meaning-based process.

- SCAPE[2] - SCAlable Preservation Environments - will address scalability of large-scale digital preservation workflows. The project aims to enhance the state of the art in three concrete and significant ways. First, it will develop infrastructure and tools for scalable preservation actions; second, it will provide a framework for automated, quality-assured preservation workflows; and, third, it will integrate these components with a policy-based preservation planning and watch system. These concrete project results will be driven by requirements from, and in turn validated within, three large-scale testbeds from diverse application areas: web content, digital repositories, and research data sets.

[1] `http://www.arcomem.eu/`
[2] `http://www.scape-project.eu/`

- ENSURE[3] - Starting with the philosophy that "one size does *not* fit all", ENSURE (Enabling kNowledge Sustainability, Usability and Recovery for Economic value) is building on existing tools, processes and approaches to create a flexible, self-configuring software stack. The solution stack will pick both the configuration and preservation lifecycle processes in order to create a financially viable solution for the given preservation requirements, trading off the cost of preservation against the value of the preserved data over time. The requirements and validation of ENSURE are driven by health-care, clinical trials, and financial use cases.

- TIMBUS[4] - Digital Preservation for Timeless Business Processes and Services. The digital preservation problem is well-understood for query-centric information scenarios but has been less explored for scenarios where the important digital information to be preserved is the execution context within which data are processed, analyzed, transformed and rendered. It is this scenario which TIMBUS addresses. The industrial case studies — addressing business processes that include sensor hardware through to large enterprise software services — focus on: (1) engineering services and systems for digital preservation; (2)civil engineering infrastructures; (3)e-science and mathematical simulations.

In rest of the paper is organized as follows: Section 2 presents the motivations of each project, followed by comparing and contrasting them. Section 3 and 4 does the same for the objectives and approaches of the projects respectively. Section 5 discuss related work. Section 6 summarizes.

2. MOTIVATIONS OF PROJECTS

2.1 ARCOMEM

The report *Sustainable Economics for a Digital Planet*[5] states that "the first challenge for preservation arises when demand is diffuse or weakly articulated". This is especially the case for non-traditional digital publications, e.g., blogs, collaborative space or digital lab books. The challenge with new forms of publications is that there can be a lack of alignment between what institutions see as worth preserving, what the owners see as a current value, and the incentive to preserve as well as the rapidness at which decisions have to be made. For ephemeral publications such as the web, this misalignment often results in irreparable loss. Given the deluge of digital information created and this situation of uncertainty, a first necessary step is to be able to respond quickly, even if preliminarily, by the timely creation of archives, with minimum overhead that would enable later engagement in more costly preservation actions. This is the challenge that ARCOMEM is addressing, relying on the "wisdom of the crowds" for intelligent content appraisal, selection, contextualization and preservation.

The Social Web not only provides a rich source of user generated content. It also contextualizes content and reflects content understanding and appraisal within society. This is done by interlinking, discussing, commenting, rating, referencing, and re-using content. The ARCOMEM project will analyze and mine this rich social tapestry to find clues for deciding what should be preserved (based on its reflection in the Social Web), to contextualize content within digital archives based on their Social Web context, and determine how to to best preserve this context. The Social Web based contextualization will be complemented by exploring topic-centered, event-centered and entity-centred processes for content appraisal and acquisition as well as rich preservation.

Two application scenarios are used for validating and showcasing the ARCOMEM technology. The first application will target the Social Web driven event and entity aware enrichment of media-related Web archives as they are, for example, required by broadcasting companies. This showcase will be driven by the broadcasting companies Sudwestrundfunk (SWR) and Deutsche Welle. The second ARCOMEM application will validate and showcase the use of ARCOMEM technology for the effective creation of Social-web-aware political archives based on Web archives and other digital archives. This will be driven by the Hellenic Austrian Parliaments.

2.2 SCAPE

The fact that the volume of digital content worldwide is increasing geometrically demands that preservation activities become more scalable. The economics of long-term storage and access demand that they become more automated. Unfortunately, the present state of the art fails to address the need for scalable automated solutions for tasks like the characterization or migration of very large collections. Standard tools break down when faced with very large or complex digital objects; standard workflows break down when faced with a very large number of objects or heterogeneous collections. Even the preservation systems used in the largest memory institutions lack the necessary automated quality assurance tools for detecting and reporting errors in a preservation process, and thereby fail to fully mitigate preservation risks.

SCAPE will use these large testbeds to define its requirements and validate its results. The **Web Content Testbed** highlights the challenges presented by heterogeneous collections and a rapidly changing delivery environment. The sheer volume of content in web archives requires fully automated, scalable preservation solutions. Web content covers many diverse file formats in multiple versions, including obsolete formats, and also associated rendering tools. The **Digital Repositories Testbed** highlights the challenge of carrying out preservation actions within an institutional context where there are legal and policy requirements and substantial investments in legacy systems. Preservation challenges coming from large scale digital repositories include issues of scalability along several dimensions: number, size, and complexity of digital objects as well as heterogeneity of collections. Furthermore, collection profiling is an integral part of planning. Finally, the current generation of digital library and preservation environments are often based on network service oriented architectures that do not scale. The **Research Data Sets Testbed** is concerned with the pressing need for preservation in scientific communities in the face of threats to long-term access and usability of scientific data. Particular aspects of this testbed include potentially very large data sets, wide variety of practices and unique requirements to preserve the original context of the experiment which generated the data in the first place.

[3] http://ensure-fp7.eu/
[4] http://www.timbusproject.net/

2.3 ENSURE

As opposed to the preservation of cultural heritage information, which it is presumed needs to be retained forever, it is neither economically viable, nor in some cases even desirable, to preserve all data managed by business forever. The value of such data tends to decrease over time, although on the other hand, there may be legal or regulatory reasons why aged data must still be retained.

In addition to the business value of data changing over time, the appropriateness of an originally chosen preservation solution can be affected over time by changing regulations, or new advances in underlying technologies influencing price-driven solutions. ENSURE is researching how current Lifecycle Management tools can be used to control the *preservation lifecycle* amidst these shifting conditions.

In addition to examining trading off cost versus quality, ENSURE is looking into the use of emerging ICT technologies to enable solutions which are not only economical, but also capable of scaling over time to meet ever expanding amounts of data. Cloud storage is seen as a primary candidate for the underlying storage services, but this introduces additional challenges, e.g., the migration of data from cloud to cloud, security issues, and the ability to perform preservation-related computing near the storage.

The ENSURE solution will be motivated and validated by three real world use cases. Selected for their relevance to data preservation, the requirements elicited by these use cases cover a wide spectrum of topics, ranging from maintaining data privacy over time, evolving ontologies, and being able to view data stored in proprietary formats decades from now. More specifically, the use cases for ENSURE are:

- Healthcare - where enormous quantities of scientific data are tied to individuals, but managed by an organization controlled by strong regulations for privacy and traceability.
- Clinical Trials - where data has both scientific and business value with strong regularity restrictions, requiring special concern for patient privacy issues.
- Financial Services - which emphasizes the long term retention of data after the regulations mandated period only as long as it has business value.

2.4 TIMBUS

A primary motivation for TIMBUS is the declining popularity of centralized, in-house business processes maintained and owned by single entities. The presence of Software as a Service (SaaS) and Internet of Services (IoS) means business processes are increasingly supported by service oriented systems where numerous services, provided by different providers, located in different geographical locations are composed to form value-added service compositions and service systems which will continue changing and evolving. Besides the advantages of SaaS and IoS, there is the danger of services and service providers disappearing (for various reasons), leaving partially complete business processes.

TIMBUS endeavors to enlarge the understanding of digital preservation to include the set of activities, processes and tools that ensure continued access to services and software necessary to produce the context within which information can be accessed, properly rendered, validated and transformed into context based knowledge. This enlarged understanding brings DP clearly into the domain of Business Continuity Management (BCM). BCM, as standardized by the British Standards Institution (BSI), is defined as:

> A holistic management process that identifies potential threats to an organization and the impacts to business operations that those threats, if realized, might cause, and which provides a framework for building organizational resilience with the capability for an effective response that safeguards the interests of its key stakeholders, reputation, brand and value-creating activities. [8]

2.5 Comparison and Contrast

Obviously, within the context of the EU call, each project has digital preservation as a motivation. However, ACROMEM stands alone in dealing with publically available and non-regulated data and in harnessing the "Wisdom of the Crowds" to help decide what to preserve. TIMBUS focuses on the environments that produce the data rather than the data itself. ENSURE and TIMBUS are motivated in part by accurate risk assessment and preservation lifecycle issues related to regulations. Together with SCAPE, they also address the scalability of technology infrastructure and software infrastructure for digital preservation. While there is some overlap in use cases,the projects as a whole cover a broad cross section of scenarios from tradition memory institutions (SCAPE), web (SCAPE, ACROMEM), engineering (TIMBUS), scientific (SCAPE, ENSURE, TIMBUS), health care (ENSURE), and finance (ENSURE).

3. OBJECTIVES OF PROJECTS

3.1 ARCOMEM

ARCOMEM's goal is to develop methods and tools for transforming digital archives into community memories based on novel socially-aware and socially-driven preservation models. This will be done (a) by leveraging the "Wisdom of the Crowds" reflected in the rich context and reflective information in the Social Web for driving innovative, concise and socially-aware content appraisal and selection processes for preservation, taking events, entities and topics as seeds, and by encapsulating this functionality into an adaptive decision support tool for the archivist, and (b) and by using Social Web contextualization, as well as extracted information on events, topics, and entities for creating richer and socially contextualized digital archives.

To achieve its goal, the ARCOMEM project will pursue the following scientific and technological objectives.

1. **Social Web analysis and Web mining**: effective methods for the analysis of Social Web content, analysis of community structures, discovery of evidence for content appraisal, analysis of trust and provenance, and scalability of analysis methods;
2. **Event detection and consolidation**: information extraction technologies for detection of events and related entities; methods for consolidating event, entity and topic information within and between archives; models for events, covering different levels of granularity, and their relations;
3. **Perspective, opinion, and sentiment detection**: scalable methods for detecting and analyzing opinions,

perspectives taken, and sentiments expressed in the Web and especially Social Web content;

4. **Concise content purging**: detection of duplicates and near-duplicates and an adequate reflection of content diversity with respect to textual content, images, and opinions.

5. **Intelligent adaptive decision support**: methods for combining and reasoning about input from Social Web analysis, diversity and coverage, extracted information, domain knowledge, and heuristics, etc.; methods for adapting the decision strategies to inputs received;

6. **Advanced Web crawling**: the integration of event-centric and entity-centric strategies, the use of Social Web clues in crawling decisions and methods for crawling by example and integrating descriptive crawling specifications into crawling strategies;

7. **Approaches for "semantic preservation"**: methods for enabling long-term interpretability of the archive content; methods for preserving the original context of perception and discourse in a semantic way; methods for dealing with evolution on the semantic layer.

3.2 SCAPE

Based on the challenges confronting its stakeholders, the scientific and technical objectives of the SCAPE project are:

1. **Scalability**. SCAPE will address scalability in four dimensions: number of objects, size of objects, complexity of objects, and heterogeneity of collections. The project is concerned with extending repository software to store, manage, and manipulate larger objects (e.g., multi-gigabyte video streams) and a larger numbers of objects (hundreds of millions). SCAPE will also improve the ability of existing preservation tools to manage a variety of container objects and to recognize diverse object formats.

2. **Automation**. Automated workflows are state of the art; SCAPE aims to make these workflows scalable. SCAPE preservation workflows will be simple to design, making use of the well-known Taverna [19] workbench, and will be deployable and executable on large computational clusters. Automated workflows for quality assurance will be developed to accompany the preservation workflows. The project also intends to introduce automation and scalability to the areas of technology watch and preservation planning.

3. **Planning**. SCAPE will build on the award-winning preservation planning tool Plato in order to enable institutions to answer core preservation planning questions. For large heterogeneous collections, the planning tool should enable a curator to determine what tools and technologies are optimal for preservation within in a given context, defined by institutional policies. SCAPE will also advance the state of the art by delivering a catalogue of generic policy elements and a semantic representation of these elements in a machine-understandable form that can be leveraged by the planning and watch components, enabling automated policy-driven planning.

4. **Context**. In the area of research data sets, SCAPE aims to provide a methodology and tools for capturing contextual information across the entire digital object lifecycle. The advance proposed by SCAPE is to embed migration of scientific data as a preservation action in the workflow, whilst preserving the wider context in order to maintain the reusability of the data. Additional research will be dedicated towards the preservation of software. Software can be seen both as part of the representation information for the scientific data itself, but also requires preservation in its own right.

5. **Prototype**. An important goal of SCAPE is to produce a robust integrated preservation system prototype within the time-frame of the project. This prototype will be made available as open source software. SCAPE technologies are expected to be in productive use in partner institutions by the end of the project. SCAPE components should also be integrated in products offered by the project's commercial partners.

3.3 ENSURE

To meet the challenges that ENSURE addresses, four main scientific and technical objectives have been defined:

1. **Evaluate cost and value**. The value of data over time differs between different organizations and industries. While the design plans for a radio built with vacuum tubes from the 1940's may not have a high business value today, the design plans for the B52 aircraft from the same period, and still in service today, do. As the business value of data goes down, the investment that an organization is willing to make to preserve the data will similarly decrease. Defining the *quality* of preservation as inversely proportional to the risk of losing data, ENSURE will look at ways of balancing the quality of a preservation solution against its cost and the value of data over time. ENSURE will also examine how a configured solution should evolve as the cost of its underlying infrastructure changes.

2. **Preservation Lifecycle Management for different types of data**. Many organizations today manage their data with Information Lifecycle (ILM) tools. ENSURE will research the suitability of adapting today's lifecyle management tools to long term preservation. In particular, while nearly all of today's ILM tools are passive, being driven by other systems and decisions, ENSURE will create a Preservation Lifecycle Management engine which can dynamically react to triggers generated by events affecting the original preservation conditions, such as new regulations, format changes, economic changes, etc.

3. **Content-aware, long term data protection**. For a preservation solution to be acceptable, it must control access to sensitive data and prevent its leakage over time, even though the identities of users, the value of the information, the roles which can access the information, etc. may change. The definition of what constitutes Personally Identifiable Information (PII) may also evolve over time, causing previously valid assumptions of data anomymization to be violated. Additionally, a solution to these issues must scale with the

size of the preservation system, and work environments such as cloud based data storage.

4. **Scalability by leveraging wider ICT innovations**. Cloud Storage and standard virtualization technologies are promising technologies to meet the challenge of building a preservation environment which can expand over time, without having to make large capital expenditures, or encounter spiraling costs for operating expenses. However, today's storage clouds typically aim at providing low cost storage and give few guarantees to the reliability and security of the stored information. A major challenge for ENSURE is demonstrating how a preservation system can be based on such a platform.

3.4 TIMBUS

To support the continuity of business processes, TIMBUS has a number of objectives best viewed from its three stages of digital preservation effort:

1. **Expediency** of digital preservation effort - establishing the risk of not preserving and the feasibility of digitally preserving business processes. Fundamental to determining what should be preserved is analyzing the risk experienced by an organization. Analyzing risk is a complex process requiring many sources of information to be collated and reasoned over. TIMBUS will develop methods and tools that provide an itelligent enterprise risk management (iERM) approach that will support decisions relating to (1) when to preserve, (2) what to preserve and (3) how to maintain and test what has been preserved.

2. **Execution** of digital preservation process - performing the digital preservation of business processes. After the expediency has been established it is necessary to actually execute the digital preservation process. TIMBUS, will address legalities lifecycle management (LLM) and uncover the current legal issues around digitally preserving interdependent services comprising a business process.

   Today's services are deployed on multi-tier service platforms that are not engineered specifically with digital preservation in mind. TIMBUS will address reengineering existing services for digital preservation (DP) and engineering new services for digital preservation. TIMBUS will also develop verification methods for the digitally preserved business processes which will prove the current preserved business process is valid (to some preservation guarantee level) and also provide some validation of the preserved business process in the (simulated) future. Appropriately, TIMBUS will develop processes for digital preservation of business processes which will be domain specific according to the use cases and processes that are generic for adaptation to new domains. These processes will be aligned with existing digital preservation standards and be the foundation for new standards specifically designed for digital preservation of business processes.

3. **Exhumation** of digitally preserved assets - re-running a digitally preserved business process. It must be possible to exhume and rerun the preserved business process. This issue will be dealt with by the visualization and storage innovations. However, it may still be the case that periodic business process exhumation will be required to provide ongoing guarantees of integrity. Obviously the future cannot be experienced now but TIMBUS must provide some level of assurance that a digitally preserved business process can be exhumed and re-run or exhumed and integrated into future business processes. TIMBUS will simulate technology changes to help indicate process exhumation and integration is feasible.

3.5 Comparison and Contrast

Out of the four projects examined here, three of them (ENSURE, SCAPE, TIMBUS) are organization-focused concerned with preserving in-house information, whereas ARCOMEM's domain is the web. It is therefore no surprise that the objectives for the first three projects tend to be more similar than those for ARCOMEM.

Central to all of the stated preservation projects is the ability to define what data needs to be preserved. ARCOMEM, concerned with preserving content found on the Web, will be looking for how to do this by attempting to analyze the information itself in the context of the Social Web. Amongst the other three projects, both SCAPE and TIMBUS will use tools to help the person responsible for preservation decide what needs to be preserved, whereas ENSURE assumes a set of supplied business rules will give this information. It is interesting to note that TIMBUS's evaluation of what to preserve is driven by the risk of *not* saving information, whereas in ENSURE, while abiding by regulatory constraints, attempts to balance cost versus. In all cases it is recognized that human intervention will be required to come up with the final decision on what to preserve.

Scalability is an issue of concentration in ENSURE, SCAPE and ARCOMEM, although the projects are emphasizing different aspects: ENSURE will tackle scalability in terms of infrastructure support,e.g., supplying a cloud based storage back-end that can support massive preservation; SCAPE focuses supporting a large number of different objects and object types; and ARCOMEM needs to analyze huge amounts of Web content for the content selection and appraisal.

The ability to rerun software after an extended period of time is a focus of the projects, and the use of virtualization technologies is a stated goal of ENSURE and TIMBUS.

The automation of the preservation lifecycle is being dealt with by all of the organization-focused projects. While SCAPE will be creating preservation lifecycles for deployment on large computational clusters, ENSURE and TIMBUS will examine extending existing lifecycle management tools to meet the additional requirements that digital preservation entails. TIMBUS will preserve processes encoded in a lifecycle management tool, while ENSURE, like SCAPE, focuses on the lifecycle management of the preservation process itself.

Additionally, all of the organization-focused projects are concerned with automatic verification of the quality of their runtime solutions. Quality will not only be monitored as part of the preservation lifecycle by all three, but also taken into consideration in the preservation planning stage.

4. APPROACH OF PROJECTS

4.1 ARCOMEM

The envisioned ARCOMEM system is built around two

loops: content selection and content enrichment. The *content selection loop* aims at content filtering based on community reflection and appraisal. Social Web content will be analyses regarding the interlinking, the context and the popularity of web content, regarding events, topics and entities. These results are used for building the seed lists to be used by existing Web crawlers.

Within the *content enrichment loop*, newly crawled pages will be analyzed regarding topics, entities, events, perspectives, Social Web context and evolutionary aspects in order to link them to each other by the relationship between events as well as by the involved entities such as persons, organizations, locations and artifacts.

The implementation of the ARCOMEM system is structured into three main research areas. *Social Web-based content appraisal and archive contextualization* aims at the development of methods to analyze the Social Web for getting clues for content appraisal and for extracting information for the archive enrichment. Networks and media are part of a dynamic social process, rather than collections of documents; networks, contexts and meanings co-evolve. To achieve a better understanding of this process for preservation, we need to answer several questions, such as: how do we appraise and rank content in multiple forms and from multiple sources, taking into account the wealth of socially-generated information about the content itself; how is reputation built, who are the leaders and who the followers. etc.

Events, Perspectives, Topics, & their Dynamics aims at extracting information from crawled data in order to provide semantically rich metadata for organizing and contextualizing the archived collection, and for supporting intelligent and efficient crawling strategies. Content perception and memorization are typically focused on, and organized around, events, entities and/or topics. Therefore, these will also be the main ingredients for the semantic enrichment layer for transforming long-term archives into community memories.

Intelligent and Collaborative Content Acquisition Support will focus on intelligent, adaptive and collaborative methods for driving and prioritizing the content acquisition and curation process. The main outcome comprises a prioritized list of sources to be crawled. This decision is primarily based on the relevance, importance, coverage and diversity of the content. This is complemented with an adaptation process involving the archivist or other archive users, and support the collaborative creation and management of archives by communities of curators.

4.2 SCAPE

The approach of SCAPE is dictated by four research and development sub-projects: Testbeds, Preservation Components, Platform, and Planning and Watch.

The Testbeds sub-project is the primary driver of the rest of the project in that it determines the use case scenarios, defines the preservation workflows, and evaluates the platform. The main goal is to assess the large scale applicability of the SCAPE Preservation Platform and the preservation components developed within the project. Using these software components, it creates test environments for the different application scenarios and complex large scale preservation workflows. As part of the testbed evaluation methodology, the automated planning tool will be used to evaluate the strengths and weaknesses of the action components in several scenarios.

The Preservation Components sub-project should address three known limitations of the functional components of a digital preservation system namely scalability, functional coverage, and quality. This sub-project will improve and extend existing tools, develop new ones where necessary, and apply proven approaches like image and patterns analysis to the problem of ensuring quality in digital preservation. Building on the state of the art and focusing on formats and tools that are considered most important by the Testbed sub-project, SCAPE will investigate methods to parallelize and embed components in robust and scalable workflows. SCAPE will provide the ability to capture relevant provenance and contextual information and metadata, as well as the ability to provide usable outputs for automated policy-driven preservation. Finally, SCAPE will develop new methods to automatically detect quality defaults, based on conversion of objects into images to apply image analysis techniques to detect differences resulting from preservation actions.

The SCAPE Preservation Platform will provide an extensible infrastructure for the execution of digital preservation processes on large volumes of data. The Platform sub-project will provide a flexible mechanism for the integration of existing digital repository systems and provide a reference implementation. The Preservation Platform will also provide the underlying runtime environment for large-scale testing and evaluation performed within the Testbed and Planning and Watch sub-projects. The computational layer of the Preservation Platform system will make use of Hadoop, an open-source map/reduce engine, and the underlying distributed storage layer will be based on HBase, which provides high performance and scalable data storage on top of Hadoop's Distributed File System (HDFS) [6].

The sub-project Planning and Watch addresses the bottleneck of decision processes and processing information required for decision making. This sub-project will begin with a conceptual analysis based on extensive real-world application experience. It will also define and model a set of essential policy elements in order to create a policy catalog. In the implementation phase, the machine-understandable policy representation will feed into the first release of the automated planning component. Building on this, the core watch services will be delivered. These services will in part be based on the analysis of file-type trends in web harvests. In the final phase the policy-aware planning component will be fully integrated with the platform and repository operations.

4.3 ENSURE

ENSURE's architecture consists of:

- a set of plug-ins that provide specific functionality such as format management, regulatory compliance, integrity checks, access to specific storage clouds etc.
- a runtime SOS framework that allows composing an OAIS solution [30] from appropriate plug-ins to meet a user's requirements including economic considerations,
- a configurator and a cost/performance/quality analysis engine witch can evaluate a proposed preservation solution

The *ENSURE Configuration Layer* runs prior to the initial deployment of the solution and re-executes periodically or if there are major environmental changes. Based upon the external requirements and observations on changes to the environment, the configurator can propose several possible solutions. These solutions are composed by choosing a set of plug-ins for the ENSURE framework, which, when taken together, meet the requirements. These candidate solutions are then evaluated and optimized by cost and performance models and evaluated by the preservation planning layer determining the quality of the proposed solution. Based upon this analysis, an administrator can choose the appropriate solution to deploy.

The second major layer containing our innovations is the *ENSURE Preservation Runtime*. The runtime layer is the SOA infrastructure for executing the plug-ins selected by the configuration layer. This layer provides data management and archival storage as well as ingest and access. In addition, this layer interacts with external storage services which provide the physical space for storing the preserved object and which may provide mechanisms for offloading certain preservation-related computations to be "closer" to the objects.

The ENSURE Preservation Runtime layer has four components:

- *Preservation Digital Asset Lifecycle Management* that manages the workflow of the information being preserved, from the time it is handed over to the system until the time it can be deleted since it is no longer needed. This component provides the glue for invoking the other components in the system and provides search capabilities based on ontology evolution.

- *Content-Aware Long-Term Data Protection* is responsible for the long term protection of the digital information, managing changes in what it means to secure information over time. ENSURE will focus on long term access control, long term privacy via the use of appropriate de-identification mechanisms, and intellectual property protection.

- *Preservation Runtime Infrastructure* will support a range of approaches to future accessibility including both transformation and virtualization

- *Preservation-aware Storage Services* provides the interface and mechanisms that enable storing the digital resources managed by the preservation solution in external storage services, such as clouds, and implementing preservation actions, such as integrity checks, near the data.

4.4 TIMBUS

As stated previously, TIMBUS views the digital preservation of business processes as three stages:

1. ***Expediency*** *of digital preservation effort.*

   A crucial aspect of enterprise risk management with regard to digital preservation of business processes is a careful analysis of the service dependencies in a specific business process. The following are some of the common types of dependencies that need to be preserved:

   - A *needs* B — A can only be made available when B has previously been available. For A to be preserved, B must be preserved.
   - A *substitutes* B — A can be used as a replacement for B. A can be preserved instead of B.
   - A *mirrors B* — the behavior and data of A must maintain consistency with the behaviour and data of B. A load-balancing capability and availability property must be preserved.

   A service is preserved if and only if there is some assurance that its complete dependency graph can be reconstructed at any lifetime t, where 0 <t <= PG, and PG is the *Preservation Guarantee* provided.

2. ***Execution*** *of digital preservation process.* When preserving business processes comprised of many interconnected services the legal/regulatory issues become more difficult to maintain and evolve over a long period of time. Legalities Lifecycle Management (LLM) consists of four parts: (1) intellectual property management; (2) IT contracting; (3) data protection; (4) monitoring of legal obligations to preserve. TIMBUS will develop innovative legal/regulatory processes and tools that could be incorporated into commercial ILM products. The tools will be *aware* of legal issues and also changes to legal issues or the introduction of new regulatory standards.

   Digitally preserving a business process that may be comprised of hardware devices and multi–tier service platforms will be easiest if all services are specifically engineered for preservation. However, it is also vital to address the current situation, i.e., services not engineered for preservation. TIMBUS will approach both tasks by focusing on the interfaces and metadata produced by services and the producing/consuming mechanisms.

   Server and desktop virtualization is one of the more significant technologies to impact computing in the last few years. Using virtualisation technology, a business process of distributed inter-dependent services can operate as one "virtual" system. The convergence of affordable, powerful platforms and robust scalable virtualization solutions is spurring many technologists to examine the broad range of uses for virtualisation. For very long life cycles it may also be necessary to provide support for stacked virtualisation (when support for a virtualisation technology ends and the virtualised business process needs to be virtualised again).

   Storage of the digitally preserved business process will also be an issue. Should the business process be stored as one large object? Should it be stored as a set of virtualised inter-dependent services? Should it be stored by an independent storage provider? Can it be stored by a group and spread across different locations? TIMBUS will work with the use case partners to establish a set of business process storage models that are informed by legalities/regulations, security, integrity, and so forth.

3. ***Exhumation*** *of digitally preserved assets.* As previously noted, we cannot go into the future to perform the rerun/integration. However, we can begin

to provide some level of *simulated future*. Our objectives in TIMBUS are: (1) exhuming the business process with the underlying infrastructure hidden – end user perspective; (2) exhuming the business process with the underlying infrastructure exposed – verification perspective; (3) exposure of appropriate metadata regarding business process and supporting software/technology stack; (4) interfacing with other services via standardized information exchange formats specifically addressing digital preservation concerns; (5) a "future simulated" test bed providing guarantee of the preserved business process rerunning and integrating by simulating future changes such as new file formats, interface changes, OS changes, storage changes, database changes, etc.

4.5 Comparison and Contrast

All four projects intend to develop prototype software frameworks. SCAPE, ENSURE, and TIMBUS propose to implement platforms for the execution of preservation processes or workflows. Both SCAPE and ENSURE propose service-oriented architectures (SOA), although SCAPE intends to use SOA workflows as prototypes that should later be executed on a parallel processing architecture. TIMBUS is concentrating on the legal and IPR aspects of the digital lifecycle, while ENSURE is more concerned with economic cost/quality/performance trade-offs and how these are managed as part of information lifecycle management. ARCOMEM's two stage workflow (content selection, content enrichment) is, in contrast, highly specialized for the web archiving use case.

Both SCAPE and ARCOMEM hope to use the Internet itself as a guide for preservation practices. In the case of ARCOMEM the content of social media should guide the harvesting process; in the case of SCAPE, trends that can be observed from Internet harvesting (for example, the frequency distribution of file types) will be used as input for the automated preservation planning process. ENSURE foresees a configuration layer that manages preservation planning, again with specific emphasis on cost, performance and quality trade-offs. TIMBUS proposes a unique approach to planning through dependency and risk management.

Both ENSURE and TIMBUS explicitly plan to use virtualisation as a tool for preservation, although ENSURE appears to focus more on using virtualisation and emulation in order to access digital objects, whereas TIMBUS sees virtualisation as a means to preserve and recover entire business processes.

5. RELATED WORK

Because the projects we describe touch on so many aspects of digital preservation, there is a broad set of related work. Clearly these projects all build on prior major efforts such as CASPAR [9] and Planets [32] and standards such as OAIS [30]. And while there is overlap in the relevant prior art, each project pulls in its own specific related work. Given the breadth of areas touched, this description of related work only scratches the surface.

Service Oriented Architecture (SOA)/Service Oriented Computing is relevant to SCAPE, TIMBUS and ENSURE. Many prior preservation approaches, e.g., [32, 9], built on SOA. While SOAs have many positive aspects, based upon the Planets' experience, SCAPE concluded that there is a need for a more scalable approach to processing the vast amounts of data managed in a large scale digital preservation solution. One specific concern is the difficulty of defining, debugging and executing complex preservation flows in a SOA framework. Another concern is the overhead both on the network and computation in processing the text intensive SOA protocol.

Grid infrastructures address some of these concerns. Data grids, such as Integrated Rule-Oriented Data System (iRODS) [21] can manage huge amounts of scientific data dispersed over heterogeneous sites. As depicted in [18], it is conceptually possible to model simple workflows using the iRODS's rule declaration language. The relative complexity of the iRODS technical language, however, makes it inappropriate for use by workflow designers; on the other hand, it is possible to use a workflow engine like Taverna [19] on top of iRODS storage layer. Even if we use a tool such as Taverna to define the workflows, we still need to consider that data-grid approaches primarily focus on data access, replication, and bit-stream integrity rather than providing data-intensive execution capabilities.

One paradigm for executing operations in parallel is the cloud-derived MapReduce framework [13]. It provides an abstraction for a highly parallel data flow architecture where each processing step operates on some partition of the very large data set. Hadoop [4] is a publicly available MapReduce engine. SCAPE and ARCOMEM will build on efforts like Hadoop to address the research challenges outlined above, processing large numbers of objects in parallel. Initial experiments have already demonstrated the feasibility of this approach [36].

Related to this use of cloud-derived technologies for scale-out computation is ENSURE's use of storage clouds (public or private) for digital preservation. Storage clouds, with their pay-per-use model, are one of the most important new ICT trends. However, the immaturity of these offerings leads to questions on their appropriateness for digital preservation [34, 25]. In spite of these concerns, there have been initial efforts to use storage clouds as the infrastructure for digital preservation. Most notable is DuraCloud [14] which offers a service that can run on multiple cloud providers and which provides the first strong example of building preservation solutions on clouds, addressing issues such as using a cloud for a backup copy,working with multiple cloud providers and running compute jobs on the preserved content in the cloud. ENSURE will build on the concepts and approach of DuraCloud, examining ways to address the concerns that exist in using a storage cloud for preservation. In particular ENSURE will look at how to integrate into a preservation solution, concepts such as as proofs of retrievability/data possession [7, 12] and provenance tracking in the cloud [29]. To enable scalability, like SCAPE, ENSURE will examine how to move preservation computation closer to the data, building on CASPAR's Preservation DataStores (PDS) [33] and emerging paradigms for compute near storage such as the aforementioned MapReduce.

The preservation of service-based processes becomes a challenge of scale. Unlike static software components, for which preservation approaches exist, e.g., emulation and versioning solutions, service-based processes are characterized by being dynamic, frequent reconfigurations, replacement of single components, continuous release cycles, and dependency on informal contextual parameters. This makes it hard to

always have a complete snapshot of an entire system/process to preserve. Learning and reasoning techniques have to be employed and handle changes.

With regard to TIMBUS, the EDOS [15] EU project provides techniques and tools for quality assurance and better dependency management of service based processes. The MANCOOSI [26] EU research project is also relevant to TIMBUS. It is encodes the relationships between software components such as dependencies and conflicts, and solves dependencies encoded as a multi-criterion optimization problem with different utility functions, e.g., cost of the software, time to setup, and human resources, etc.

One area of focus for TIMBUS is Intelligent Enterprise Risk Management. Understanding enterprise risk, not just financial risk, has been addressed from a business continuity management perspective. Approaches to model and analyse resource dependencies, failure propagation and recovery models will be used as a baseline and include Failure Mode, Effects, and Criticality Analysis (FMECA) [17, 31], Fault Tree Analysis (FTA) [11], Tropos Goal-Risk Framework [3], Risk Aware Process Evaluation (ROPE) [39].

Related to this analysis of risk in TIMBUS, ENSURE examines cost value trade-offs. ENSURE is not the first project to consider the cost of digital preservation. For instance, [16] and [37], among others, both describe approaches to modelling the cost of a preservation solution. There is also work to evaluate the quality of solutions, with tools such as [24, 20] which build on the emerging ISO standard for Audit and Certification of Trustworthy Digital Repositories. ENSURE goes beyond these efforts by adapting techniques such as benchmarking models [10] and extends these approaches with a view of whole life cycle cost to address obsolescence [35, 38] to allow cost/value tradeoffs.

Protecting data over the long term, which is one of the focus areas of ENSURE, has multiple aspects. One of the more significant is to prevent leakage of personally identifiable information. De-identification is a common approach to facilitate secondary use of personal data by sanitizing the data. Beyond basic techniques which remove or mask direct identifiers, more advanced techniques, e.g., [22], address more sophisticated re-identification attacks. None of these approaches, however, address the fact that what constitutes personally identifiable information changes over time.

Several projects have pursued Web archiving (e.g., [2, 1]). The Heritrix crawler [28], jointly developed by several Scandinavian national libraries and the Internet Archive through the International Internet Preservation Consortium (IIPC), is a mature and efficient tool for large-scale, archival-quality crawling. The IIPC has also developed or sponsored the development of additional open-source tools and an ISO standard for web archives (ISO 28500 WARC). On the operational side, the Internet Archive and its European sibling, the Internet Memory Foundation, have compiled a repository of more than 1 Petabyte of web content which is growing at 100 Terabytes per year. A large number of national libraries and national archives are now also actively archiving the Web as part of their heritage preservation mission.

The method of choice for memory institutions is client-side archiving based on crawling. This method is derived from search engine crawl, and has been evolved by the archiving community to achieve a better completeness of capture and to increase temporal coherence of crawls. These two requirements (completeness and temporal coherence) come from the fact that, for web archiving, crawlers are used to build collections and not only index [27]. These issues were addressed by LiWA (Living Web Archives) [23], which also develops new approaches for the capturing of rich and complex web content, data cleansing and filtering, and archive interpretability.

6. SUMMARY

We presented the four new large digital preservation projects funded by the EC that started in 2011: ACROMEM, SCAPE, ENSURE, and TIMBUS. The motivation for all projects is expanding the scope of long term digital preservation. However the use cases motivating the work vary from publicly available data on the web to data of commercial organizations. The data spans beyond documents to commercial, medical, and scientific data that needs to be interpreted by programs or workflows.

The objectives of the projects spans from methods to define what should be preserved to building the preservation environment. For deciding what to preserve different methods are planned, varying from use of social web, to risk and cost based approaches, and considerations of data protection. For preservation environments, scalability, reconfigurability, supporting different types of data, supporting preservation software, and handling the full lifecycle of preservation are among the areas addressed by the projects. All the four projects plan to develop prototype tools and to build on results of previous projects.

While the projects presented here differ in their objectives and approaches, together they try to cover a bigger part of the long term digital preservation problem by addressing wider range of organizations that need preservation, more types of data, and practical problems of tools and scalability. As the projects progress in the following years, interoperability between those projects and with other digital preservation efforts will be considered. Our hope is that our efforts will make digital preservation more accessible and will contribute to future usability of our digital information.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh FrameworkProgramme (FP7/2007-2013) under grant agreements No 270000, 270239, 270137, and 269940.

8. REFERENCES

[1] S. Abiteboul, G. Cobena, J. Masanes, and G. Sedrati. A First Experience in Archiving the French Web. In *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '02, pages 1–15, London, UK, UK, 2002. Springer-Verlag.

[2] A. Arvidson and F. Lettenström. The Kulturarw Project - The Swedish Royal Web Archive. *Electronic library*, 16(2), 1998.

[3] Y. Asnar and P. Giorgini. Modelling Risk and Identifying Countermeasure in Organizations. In J. Lopez, editor, *1st International Workshop on Critical Information Infrastructures Security*, volume 4347 of *Lecture Notes in Computer Science*, pages 55–66. Springer-Verlag, 2006.

[4] A. Bialecki, M. Cafarella, D. Cutting, and O. O'Malley. Hadoop: a framework for running applications on large clusters built of commodity hardware. *Wiki at http://lucene. apache. org/hadoop*, 2005.

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

[6] D. Borthakur. *The Hadoop Distributed File System: Architecture and Design*. The Apache Software Foundation, 2007.

[7] K. Bowers, A. Juels, and A. Oprea. HAIL: A high-availability and integrity layer for cloud storage. *ACM CCS*, November 2009.

[8] BSI. BS 25999-1:2006 Business continuity management. Code of practice., 2006.

[9] CASPAR Digital Preservation User Community. `http://www.casparpreserves.eu/`, 2010.

[10] C. Chituc and S. Nof. The Join/Leave/Remain (JLR) decision in collaborative networked organizations. *Computers & Industrial Engineering*, 53(1):173–195, 2007.

[11] I. E. Commission. IEC 61025. Fault Tree Analysis, Ed. 2.0, 2006.

[12] R. Curtmola, O. Khan, and R. Burns. Robust remote data checking. In *StorageSS '08: Proceedings of the 4th ACM international workshop on Storage security and survivability*, pages 63–68, New York, NY, USA, 2008. ACM.

[13] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. *Commun. ACM*, 51:107–113, January 2008.

[14] DuraCloud. `http://www.duraspace.org/duracloud.php`.

[15] EDOS - Environment for the development and Distribution of Open Source software. `http://www.edos-project.org`.

[16] K. Fontaine, G. Hunolt, A. Booth, and M. Banks. Observations on cost modeling and performance measurement of long term archives. In *PV Conference*, 2007.

[17] Y. Haimes. *Risk Modeling, Assessment, and Management*. John Wiley & Sons, Inc, 2009.

[18] M. Hedges, T. Blanke, and A. Hasan. Rule-based curation and preservation of data: A data grid approach using iRODS. *Future Gener. Comput. Syst.*, 25:446–452, April 2009.

[19] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34(Web Server issue):729–732, July 2006.

[20] P. Innocenti, S. Ross, E. Maceviciute, T. Wilson, J. Ludwig, and W. Pempe. Assessing digital preservation frameworks: the approach of the SHAMAN project. In *MEDES '09: Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, pages 412–416, New York, NY, USA, 2009.

[21] IRODS: Data Grids, Digital Libraries, Persistent Archives, and Real-time Data Systems. `http://irods.sdsc.edu/index.php/Main\_Page`.

[22] N. Li, T. Li, and S. Venkatasubramanian. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In *ICDE*, pages 106–115. IEEE, 2007.

[23] LiWA – Living Web Archives. `http://www.liwa-project.eu/`, 2011.

[24] Long Term Digital Preservation Assessment Tool, IBM Haifa Research Lab,. `https://www.research.ibm.com/haifa/projects/storage/datastores/ltdp.html`.

[25] Long-term Preservation Storage: OCLC Digital Archive versus Amazon S3. `http://dltj.org/article/oclc-digital-archive-vs-amazon-s3/`.

[26] MANCOOSI - managing software complexity. `http://www.mancoosi.org`.

[27] J. Masanès. *Web archiving*. Springer, 2006.

[28] G. Mohr, M. Kimpton, M. Stack, and I. Ranitovic. Introduction to heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW04)*, 2004.

[29] K.-K. Muniswamy-Reddy, P. Macko, and M. Seltzer. Provenance for the Cloud. In *8th USENIX Conference on File and Storage Technologies (FAST '10)*, San Jose, CA, USA, Feb. 2010. USENIX.

[30] *OAIS: Space data and information transfer systems – Open archival information system – Reference model*. ISO 14721:2003, 2003.

[31] I. A. Papazoglou, O. N. Aneziris, J. G. Post, and B. J. M. Ale. Technical modeling in integrated risk assessment of chemical installations. *Journal of Loss Prevention in the Process Industries*, 15(6):545 – 554, 2002.

[32] PLANETS home. `http://www.planets-project.eu/`, 2010.

[33] S. Rabinovici-Cohen, M. Factor, D. Naor, L. Ramati, P. Reshef, S. Ronen, J. Satran, and D. L. Giaretta. Preservation DataStores: new storage paradigm for preservation environments. *IBM Journal of Research and Development*, 52(4):389–399, 2008.

[34] D. Rosenthal. Preservation in the Cloud. In *Preservation in the Cloud*. Library of Congress, September 2009.

[35] P. Sandborn and G. Plunkett. The other half of the DMSMS problem-software obsolescence. *DMSMS Knowledge Sharing Portal Newsletter*, 4(4):3, 2006.

[36] R. Schmidt, C. Sadilek, and R. King. Workflow System for Data Processing on Virtual Resources. *International Journal on Advances in Software*, 2(2–3):234–244, 2009.

[37] J. Slats and R. Verdegem. SCost Model for Digital Preservation. `http://dlmforum.typepad.com/Paper\_RemcoVerdegem\_and\_JS\_CostModelfordigitalpreservation.pdf`, 2010.

[38] S. Strodl, C. Becker, R. Neumayer, and A. Rauber. How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, page 38. ACM, 2007.

[39] S. Tjoa, S. Jakoubi, and G. Quirchmayr. Enhancing Business Impact Analysis and Risk Assessment Applying a Risk-Aware Business Process Modeling and Simulation Methodology. In *ARES*, pages 179–186, 2008.

Recordkeeping in Temporary Command Settings

Erik A.M. Borglund
Mid Sweden University
Universitetsbacken 1
871 88 Härnösand, Sweden
+46 611 86233
erik.borglund@miun.se

ABSTRACT
This article is about the recordkeeping that takes place during large police operations in different command post settings, and presents the tentative results from a three-year study. The aim is to increase knowledge of the problems related to recordkeeping in this kind of environment. Two police operations have been used as data sources, one large EU ministerial meeting in Sweden, and one regional disaster training exercise. An interpretative case study approach has been applied where observation and interviews were the primary data collection techniques. One problem is that too much important information is recorded in less permanent ways, on whiteboards and on flip charts. Capture, storage, and dissemination of those temporal and analogue records are difficult, and reduce the possibility to use records from large operations as knowledge reservoirs.

Categories and Subject Descriptors
D.2.7 [**Distribution, Maintenance, and Enhancement**]: Documentation; I.7.1 [**Document and Text Editing**]: Document management

General Terms
Management, Documentation,

Keywords
Knowledge management, Preservation, Police operation, Recordkeeping.

1. INTRODUCTION
During large crisis or emergency situations it is common that the responders organize themselves in temporary organizations to manage the incident/situation. This article is about recordkeeping, and especially about the records born in these temporary established command settings, and the problems and challenges that digital preservation brings to these records and their management.

Managing records is also about managing documented knowledge, as records are sometimes referred to as organizational knowledge sources [1-4]. It is also possible to claim that records are "institutional memory" [4]. In modern recordkeeping theory, i.e. the records continuum model, records constitute the archive and the archive forms part of corporate and collective memory. Organizations can gain advantages in productivity and innovation due to knowledge management. [5]. By managing records with high efficiency and quality during a large crisis, knowledge can be gained and preserved for future use within the organization and for other organizations. According to Chua [6] disaster management has not been the traditional domain for applying knowledge management, even if the societal impact could be high. However there are a few research contributions to find [6-8].

This article aims to increase the knowledge about the challenges and problems related to recordkeeping in large-scale police operations. The results are tentative from an ongoing research project with the long-term goal to develop efficient recordkeeping strategies and methods that over time can ensure that Knowledge gained from large operations can be reused. The argument for the importance of this research is the definition of knowledge management applied in this article from Jennex "KM is the practice of selectively applying knowledge from previous experiences of decision making to current and future decision making activities with the express purpose of improving the organization's effectiveness."[9]. Without a working recordkeeping this can be difficult to achieve.

2. RESEARCH APPROACH
The applied research method is best described as an interpretative case study [10, 11]. The data collection methods have primarily been field studies with participatory observations [12], and interviews [13]. The data collection and analysis has been an effort of one researcher, which has 20 years experience from the police practice as a sworn officer and who has now become an academic. The dual role, officer and researcher, implies that the researcher can be seen as a reflective practitioner [14].

The results presented in this article are based upon empirical data and have been evolved through analysis of the collected data iteratively.

2.1 Research setting and data collection
This paper presents the tentative results from a three-year research project, and is based upon two research studies.

The first study was carried out during the informal meeting for the EU energy and environment ministers held in Åre on 23–25 July 2009.

The second research study was a large regional disaster training exercise that took place in April 2010 in the County of Jämtland. The scenario of the exercise was an airplane incident where a

iPRES2011, Nov. 1–4, 2011, Singapore.

Boeing 737 from Arlanda Airport crashed during landing at Åre-Östersund airport.

3. Results

In this section we present the result from this study. We start by presenting how the police was organized during these two police operations, and within this presentation we also present how records are managed during the police operations.

3.1 EU meeting in Åre

The police operation responsible for the security during the EU minister meeting was organized in one support of staff in Östersund, and two command posts in Åre.

3.1.1 Support of staff in Östersund

The support of staff was organized following the National standard for how large police operations should be arranged [15], and had the following competences and units:

Chief of Staff
P1 – Operational Staff
P2 – Intelligence
P3 – Operational management
P4 – Logistics and equipment
P5 – Planning and co-operation
P6 – Operational Analysis
P7 – Information
P8 – Various tasks

The support of staff had ordinary meetings at 7am, 9am, 3pm, and 10pm. During these meetings all the staff members were gathered in the staff room and each of them presented what new information had been received since the last meeting. These meetings were documented in minutes, that after the meetings were stored in a common folder on the police Intranet, which all staff members had access to. On the walls in the staff room all upcoming events were plotted together with information of common interest on whiteboards and on flip charts.

In between the staff meetings each of above mentioned P-functions (P1-P8) reported to the Chief of Staff important events. This information was communicated face to face, by telephone or by mail even if the latter was very rare.

3.1.2 Command posts

During the police operation in Åre, the police had two command posts aimed to locally manage the police operation on site. They were both located in large rooms (one for each command post) and the rooms were equipped with whiteboards and flip charts. On the flip charts and whiteboards information about e.g. important events, time schedules, and similar was written down. The command posts had fixed meeting times, where they summarized the past and informed and updated the support of staff and the police operational commanders. The information presented at these meetings were documented into minutes, and kept in a common folder on the police Intranet.

Between the meetings the officers in charge at the two command posts recorded all events in a Word document, as a diary with time stamps and a description of what happened together with the reporting officer signatures, which also was stored in the common folder. The different actors stationed at the command posts shared information between each other verbally.

In addition to the two command posts, the police operation also had a mobile command post. In the mobile command post mobile phone and radio were used as tools to communicate with the other police units. Important decisions taken by the police operation commander were documented in a log in the mobile command post, and afterwards transferred to an electronic document.

3.2 Regional disaster training exercise

The regional disaster training exercise involved only one large support of staff within the police. The support of staff was organized and physically placed in a room next to the command and control centre at the Östersund Police Authority. The support of staff was organized with a Chief of Staff, and all P-functions (P1–P8). The room was specially designed for such purposes and equipped with two dispatch operation tables, one fixed large whiteboard, one mobile whiteboard, and two video projectors (figure. 1).

During the disaster training exercise the majority of the members of the support of staff worked in this specially designed room. Only the P7 function, with responsibility for the press, moved back and forth during the disaster training exercise.

Figure 1. The Support of staff setting, with all P-functions gathered around the table

As an effect of that this was a disaster training exercise, the support of staff had rather frequent meetings, where they summarized the ongoing activities. Every function documented their work in Word files in shared folders on the police Intranet. All of the known information was plotted on whiteboards, but also on large paper sheets (from the flip chart) that were put on the walls with adhesive tape. Due to the fact that this was a regional disaster training exercise, there was also an aim to reach a wider cooperation between the police and other actors involved in the disaster training exercise, e.g. the fire brigade, the medical service, the municipalities, the air craft company etc. These cooperative meetings were held almost every hour in a joint support of staff room, where liaison officers represented all actors. These meetings were documented by the police liaison officer, and the core content of the meetings was reported during the regular support of staff meetings.

4. DISCUSSION

The operational activities, and operational decisions during the EU meeting in Åre and during the regional disaster training exercise were documented in the command and control system. The activities and decisions that took place in the temporal command settings, as support of staff and command post, mainly used Microsoft Office Word™, flip charts, and whiteboards to document their work. Almost no activities from the support of staff and the command posts were recorded in ordinary police command and control system, due to the fact that much of the activities was on strategic level and not operational. In the command and control systems, normally only operational activities is stored.

4.1 Microsoft Office Word™

During the disaster training exercise the use of Microsoft Office Word™ was very limited. It is possible that this was an effect of the fact that all police officers knew that it was only an exercise. Therefore the data presented is primary based upon data from the police operation in Åre.

Every meeting, and everything that happened during the police operation in Åre were documented in Microsoft Office Word™, and the *.doc file was stored in a common folder (accessible for all officers involved in the management of the police operation). The documents were given descriptive names and a timestamp. The common folder had sub-folders for e.g. the support of staff, command posts, general decisions, and contact information.

When you open the folder you find a list of all Word files within the folder. Both the naming of the documents and the date function in the file explorer (the Swedish police use Windows XP™) Informed each user of whether or not any new documents had been added, and which documents that it might be necessary to read.

Microsoft Office Word™ was used to document and store important information more permanently than writing on whiteboards or flip charts made possible. To be fully updated, all sub-folders had to be opened and the list of documents read one by one. In Microsoft Office Word™ a rather detailed picture of on-going activities could be stored, as opposed to the information documented on whiteboards or flip charts.

Both the support of staff and the two command posts had a 24-hours diary where they wrote down every event that happened during the police operation. The event that was documented in the diary was documented even if the event also was documented in the command and control system. During calm periods the diary were updated almost in real time, but during more critical situations the diary was not updated until the responsible police officers had the time to update it. In the support of staff the various P3 functions all had their own diaries of every event that occurred in their jurisdiction (P-3T, P-3O, and P-3K). All P-functions (P1-P8) were responsible for documenting in Microsoft Office Word™.

4.2 The flipcharts

The flip charts played a similar role both during the disaster training exercise, and during the police operation in Åre. In the support of staff rooms at least one flip chart was placed, and also in the two command posts that were used during the police operation in Åre.

During calm and normal activities the flip charts were never used. However when there was a need to write down important information of more temporal character, this was done on the flip charts. During the police operation in Åre telephone numbers, call signs on the radio and flight numbers were the most frequent information written on the flipcharts. Some of this information was either transcribed into Microsoft Office Word™, or the single chart was torn off and put on a free space on a wall with adhesive tape. In figure 1, you will see some charts added to the wall in the pictures upper right corner. When the information on the flip chart (or on the charts on the wall) was no longer important for the management of the police operation, the chart was torn off.

It was not always a one-man show to decide whether to use the flip chart or not. Often the use of the flip charts was preceded by a discussion between the involved actors to document things on the flip chart. The flip chart was not used to document all kinds of stuff. Normally the flip chart was used to document extraordinary events. This also gave the flip chart a signal value, i.e. if there was text on the flip chart, every one knew that something important had happened.

4.3 Whiteboards

It was apparent in the two case studies that whiteboards were important in large police operations. Every room used for either the support of staff or the command posts had whiteboards, which were used as operational plotting tools.

In the support of staff rooms, one whiteboard was used as a place to plot all planned events in chronological order. If there was a change in the program, the information was updated and also marked as “updated”. On another whiteboard the support of staff plotted all important decisions given either by the police operational commander or the strategic commander (which of course could also be found in the common folders in the Intranet). One whiteboard was used to plot the important events that took place, i.e. a timeline of the activities that affected the police operation. Example of this timeline is seen in figure 2.

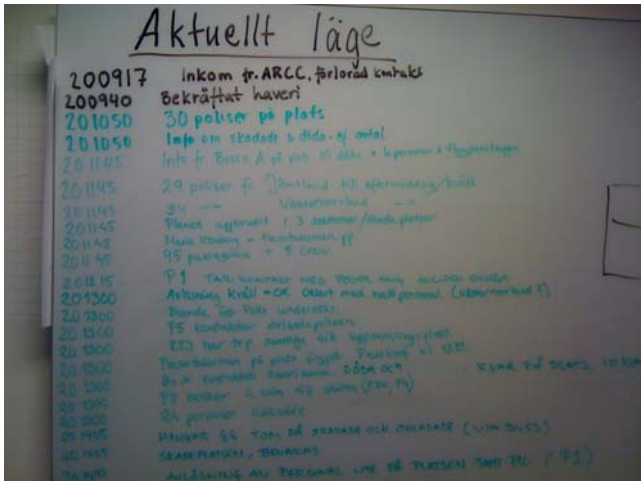


Figure 2. One whiteboard used to plot all events in chronological order

The two command posts that were set up during the police operation in Åre used boards in the same way. They used them to plot decisions, planned activities, and past activities. One

difference between the command posts and the support of staff during the police operation in Åre, was that on the whiteboards at the command posts, you also found descriptions of individuals that the police intelligence unit had analyzed and identified as potential threat to the police operation. You could also find pictures of these persons attached to the whiteboards with magnets. When a whiteboard was filled with content, as can be seen in figure 2, the police officers in the temporal command setting, either prioritized to document the content in a Microsoft Office Word™ document, or they brought in another whiteboard on which to continue writing. Sometimes they took a photo of the content.

During stressed and more chaotic situations the whiteboards played an important role in documenting ongoing activities and decisions taken. The whiteboards was easy to use and they also was accessible for all individuals in the room. Many persons could also use the whiteboard simultaneously, which made them usable during stressed situations.

4.4 Problem summary

One can easily see several problems in the way important information was recorded during the police operations used as cases in this research.

The problems are presented in relation to what can constitute the essence of document/records management, i.e. capturing, storing and dissemination of information:

Capturing. Important information is recorded on whiteboards and flip charts, which are rather temporal in their structure and implicit, which make capturing problematic. In the capturing phase we also include the capturing of metadata needed to understand the record in the future. In these two case studies both digital and analogue information is captured. This information is per definition records, as it is information that is created and maintained as some evidence over the ongoing business.

Storing. Storing is related to capturing, but is about how to store the captured record. The records that are captured in Microsoft Office Word™ have options to be stored with quality, but they are only stored on a file server in common folders. The storing of the analogue information is very ad hoc managed. Sometimes the flip chart is preserved, and sometimes the whiteboards are photographed, but this is not standard procedure.

Dissemination. During large police operations, dissemination of information is important. But dissemination of information is also important after the operation to make it possible to gain knowledge of the lessons learned from the operation. With the very analogue and temporal way to record important information during large police operations the dissemination of information both during the police operation, and after the police operation is extremely problematic. Search of recorded information is also problematic due to the same kind of arguments.

Other problems. In order to be able to evaluate large operations and in worst case investigate mistakes, traceability and accountability is two important functions that results from a qualitative recordkeeping. If knowledge from large operations should be sources for knowledge one need to further establish new methods and technologies to capture, store and disseminate the records born during these events and make sure that the necessary metadata is added. In the study no proof of a high level of knowledge amongst police officers during .

4.5 Digital preservation problem

If recordkeeping in crisis management should be an important tool for "applying knowledge from previous experiences of decision making to current and future decision making activities with the express purpose of improving the organization's effectiveness."[9], then the digital preservation is impossible to exclude. Searchability and access is not dependent upon that records are digital, but digital records can be distributed longer and accessed by actors far away. For example records created during the management of a large flooding in central Europe can be searchable and accessible for actors on another continent, who can learn from others by use the records. Therefor is it important to understand that digital preservation is embedded in modern recordkeeping. Following the record continuum model, the collective and cooperate knowledge can not be separated from the creation and capturing of a record [16].

5. CONCLUDING REMARKS

This article had an aim to increase the knowledge about the challenges and problems related to recordkeeping in large-scale police operations, and especially recordkeeping in temporal command settings.

The problems identified are all related to the analogue management of important records that are born on flip charts and whiteboards, but also to an overall lack of structure and technical support concerning document management within the police.

A continuing of this research is to study command settings where they test and evaluate more advanced technical aid, to see if the problems will be minimized or not.

6. REFERENCES

[1] Gladney, H. M. 2004. Trustworthy 100-Year Digital Objects: Evidence After Every Witness Is Dead, *ACM Transactions on Information Systems*. 22, 3, 406-436.

[2] McKemmish, S., Piggott, M., Reed, B., & Upward, F. 2005. *Archives: Recordkeeping in Society*, Wagga Wagga: Charles Sturt University, Centre for Information Studies.

[3] Menne-Haritz, A. 2001. Access-the reformulation of an archival paradigm, *Archival Science*. 1, 57-82.

[4] Sprehe, T.J. 2000. Integrating Records Management into Information Resources Management in U.S. Government Agencies. *Government Information Quarterly*. 17, 1, 13-26.

[5] Nunes, M. B., Annansingh, F., Eaglestone, B., & Wakefield, R. 2006. Knowledge management issues in knowledge-intensive SMEs. *Journal of Documentation*. 62, 1, 101-119.

[6] Chua, A.Y.K. 2007. A Tale of Two Hurricanes: Comparing Katrina and Rita Through a Knowledge Management Perspective. *Journal of the American Society for Information Science and Technology*. 58, 10, 1518-1528.

[7] Murphy, T., & Jennex, M. E. 2006. Knowledge Management, Emergency Response, and Hurricane Katrina. *International Journal of Intelligent Control And System*. 11, 4, 199-208.

[8] Chua, A. Y. K., Kaynak, S., & Foo, S.S.B. 2007. An Analysis of the Delayed Response to Hurricane Katrina Through the Lens of Knowledge Management. *Journal of The American Society for Information Science and Technology*. 58, 3, 391-403.

[9] Jennex, M. E. 2005. What is Knowledge Management?. *International Journal of Knowledge Management*. 1, 4, i-iv.

[10] Myers, M. 1997. Interpretative Research in Information Systems. In F. A. Stowell & J. Mingers (Eds.) *Information systems : an emerging discipline?* 239-266. London: McGraw-Hill.

[11] Walsham, G. 2002. Interpretative Case Studies in IS Research: Nature and Method. In M. D. Myers & D. E. Avison (Eds.) *Qualitative research in information systems : a reader*. 101-113. London: SAGE.

[12] Taylor, S. J., & Bogdan, R. 1998. *Introduction to qualitative research methods : a guidebook and resource*. New York, N.Y.; Chichester: Wiley.

[13] Kvale, S. 1997. *Den kvalitativa forskningsintervjun*, Lund: Studentlitteratur.

[14] Schön, D. A. 1983 *The reflective practitioner : how professionals think in action*, New York: Basic Books.

[15] Nylén, L. 2006. *Operativ ledning : bedömande och beslutsfattande: lednings- och fältstaber vid särskild händelse : en handledning*, Stockholm: Rikspolisstyrelsen : CRISMART, Försvarshögskolan.

[16] Upward, F. 2000. Modeling the continuum as paradigm shift in recordkeeping and archiving processes, and beyond - a personal reflection. *Records Management Journal*. 10, 3, 115-139.

We are all Archivists: Encouraging Personal Digital Archiving and Citizen Archiving on a Community Scale

Leslie Johnston
Library of Congress
101 Independence Ave SE
Washington, DC. USA 20540-1300
011-1-202-707-2801
lesliej@loc.gov

ABSTRACT

The more we interact online and manage digital lives, the more we build a born-digital record of our personal contexts and interests. There is a clear mandate that libraries, archives and museums must assume responsibility for educating the public about strategies for personal digital archiving and personal curation, and for exploring new approaches to processing, preservation, and access. Personal digital items initially have value to the individuals who generated them, but once those items are transferred to a collecting institution they will have a collective value to society. We also have a less well-established mandate to work with those citizen archivists who are taking the initiative to save large collections of digital ephemera and the at-risk output of underserved communities. Large amounts or small, all digital materials have the potential to be vital to cultural history studies in the future.

Categories and Subject Descriptors

K.4.0 [**Computers and Society**]: General

General Terms

Documentation, Experimentation.

Keywords

Personal Digital Archiving, Citizen Archiving, Web Archiving, Digital Preservation.

1. INTRODUCTION

We are all archivists. We are all makers. We are all participants in many communities. There is a seeming increase in interest in "citizen" activities: citizen journalism, citizen science, and citizen archiving. We are all creating personal histories that increasingly include born-digital records and some type of online activities. The more we interact online and manage digital lives, the more we also seek to digitize our analog histories, as well as strive to make some sort of record of our personal contexts and interests.

iPRES2011, Nov. 1–4, 2011, Singapore.

Initiatives to provide education about support personal digital archiving best practices often dovetail with initiatives to collate or collect personal digital materials. Interest in citizen archiving has moved from the analog to the digital as the awareness of the short life spans of formats, media, and online services becomes more widespread. There is an increased engagement of individuals and communities in digital preservation: personal initiatives are growing into large-scale efforts. How should the digital preservation community encourage and incorporate these activities?

2. "CITIZEN" EFFORTS

We have all seen examples of "Citizen Journalism," from the tweeting in Iran during the 2009 elections, commuters who were quick to send email with photos to media outlets immediately following 9/11 or the 7/7 London Bombings, bloggers reporting on local news in small communities where newspapers are no longer easily sustained, or something as seemingly straightforward as the ability for anyone to comment on and initiate a discussion about published news stories online. There is an almost unprecedented involvement of the general public in the process of collecting, reporting, analyzing and disseminating news and information. [1]

We might be less informed about "Citizen Science," a category of scientific work in which individuals or networks of volunteers with little or no scientific training perform scientific research tasks such as observation, measurement, or computation. Prior to the twentieth century, from the Enlightenment through the Victorian era, most scientific endeavors were undertaken by "amateurs," both trained and self-educated. Science educators encourage people of all ages that they participate in research, be it species counts, environmental and atmospheric monitoring, astronomical observation, or volunteering on an archaeological excavation.

We are of course familiar with "Citizen Archivists," a term once used primarily to refer to amateur collectors, genealogists, and family historians who amassed physical collections of photographs, personal papers, newspapers, maps, films, recordings, etc., documenting an era, event, place, community, family, or an individual. Such collections are the core of many of our institutional collections.

But Citizen Archivists are now preserving all things digital as well, from personal histories to software to games to the web itself. And we should be encouraging this.

3. PERSONAL DIGITAL ARCHIVING

One clear topic to be addressed is personal digital archiving. As Richard Cox notes in his paper on digital curation and citizen archivists:

> "It is likely that the increasing use of digital formats will enhance interest in the preservation of personal archives and that this will strengthen the public's awareness of the importance of archives, records, and information management." … "The exciting aspect of rethinking how archivists will work in preserving personal and family archives is that it may re-open a much greater possibility for reaching the public with a clearer sense of the archival mission, an objective archivists and their professional associations have struggled to do for several decades with very mixed results. It is, however, clearly the case that the public itself is actually sowing the ground for archivists to seed." [2]

The single most extreme individual effort in personal digital archiving is certainly that of Gordon Bell and his MyLifeBits lifelogging project.[1] [3, 4, 5] After responding to a colleague's request to digitize one of his books, he started to consider what it meant to have a hybrid personal history, part paper-based and part digital. He considered the combination of a small wearable camera and the Memex proposed by Vannevar Bush [6] and began to document his days with a SenseCam[2] along with digitizing his personal archives.

For everyone else, they know they need to save their personal digital output and history and records for future generations but are wondering how to do it. Cathy Marshall described the issues succinctly:

> "… people archive their personal digital belongings by relying on a combination of benign neglect, sporadic backups, and unsystematic file replication. Even the most valuable of their digital assets -- files representing considerable investments of effort, significant emotional worth, or actual cash expenditures -- are often in danger of being lost. Distributed storage, uncontrolled accumulation of digital materials, a lack of standard curation practices, and an absence of long term retrieval capabilities all point toward an incipient digital dark age." [7]

There is a recent but growing body of research on personal digital archiving. Some of this research builds on related research in Personal Information Management (PIM) strategies, the study of personal digital information seeking and management, sometimes described as "keeping found things found." [8,9,10,11,12] Cathy Marshall, Neil Beagrie, Jeremy Leighton John and others have written extensively on personal digital archiving strategies and requirements. [13, 14, 15, 16]

In that vein, there is an equally important emerging area of study on methodologies for the acquisition and management of personal digital archives at collecting institutions, as well as on the use of digital forensics tools in the appraisal, capture, management, description, and preservation of personal digital collections. [2, 17, 18, 19, 20, 21, 22] Conferences on the issues are taking place internationally[3]. Awareness of the issue has moved out into the popular press. [23, 24, 25, 26, 27] Many institutions have instituted digital forensics programs and labs to process collections, including the British Library[4], Emory University [25], Stanford University[5], and Oxford University[6].

But what about advice for individuals? Many commercial services have launched to assist in personal archiving and to provide "legacy services" in case of incapacitation or death. [27, 28, 29] Where are collecting institutions in this realm?

A number of initiatives and organizations have released resources or tools to assist in personal digital archiving. While some issues in the file format persistence have been presented in a lightedhearted way [30], there are now a number of resources to advise both individuals and institutions on the comparative sustainability of different types of files and media storage. The Library of Congress maintains its "Sustainability of Digital Formats" site.[7] The National Library of Australia has created the prototype Mediapedia, which documents storage media.[8]

The Paradigm Project released its "Guidance for Creators of Personal Papers."[9] The SALT project[10] at Stanford has been prototyping digital self-archiving "legacy" tools for faculty. , Penn State University is developing a model for curating digital intellectual lives[11]. FamilySearch has released a guide for preserving family history records. [31]

The Library of Congress has launched two personal digital archiving initiatives. In 2010 the Library hosted its inaugural Personal Archiving Day, initially held as an ALA Preservation

[1] See http://research.microsoft.com/en-us/projects/mylifebits/.

[2] See http://research.microsoft.com/en-us/um/cambridge/projects/sensecam/ .

[3] To date, two personal archiving conferences have been held at the Internet Archive (http://www.personalarchiving.com/), one was sponsored by the Digital Lives project (http://www.bl.uk/digital-lives/conference.html), one was sponsored by the Digital Curation Centre (http://www.dcc.ac.uk/node/9219), one was hosted by the University of North Carolina, Chapel Hill (http://ils.unc.edu/callee/emanuscripts-stewardship/index.html), one on forensics was hosted by the University of Maryland (http://mith.info/forensics/), and one was hosted by the Library of Congress on home movies (http://www.centerforhomemovies.org/homemoviesummit.html) , among others.

[4] See http://britishlibrary.typepad.co.uk/digital_lives/2011/03/the-emss-lab-20.html

[5] See http://lib.stanford.edu/digital-forensics

[6] See http://www.bodleian.ox.ac.uk/beam

[7] See http://www.digitalpreservation.gov/formats/

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

[9] http://www.paradigm.ac.uk/guidanceforcreators/guidance-for-creators-of-personal-digital-archives.pdf

[10] See http://sites.google.com/site/stanfordluminaryarchives/

[11] See http://www.personal.psu.edu/esc10/blogs/E-Tech/2011/02/personal-digital-archiving-201.html

Week event. Library staff members were available to answer questions about physical and digital formats, ranging from the care of daguerreotypes to the digitization of audio and video tapes to the preservation of digital images. The first event was so popular that it was repeated at the 2010 National Book Festival, and again during the 2011 ALA Preservation Week. At every event Library staff members are asked for more extensive guidance, showing that there is a huge demand for advice in this area.

To accompany the events, the Library launched its "Personal Archiving: Preserving Your Digital Memories" site, providing downloadable and printable information sheets with tips for various formats.[12] The personal archiving pages are now among the most popular on the digitalpreservation.gov site.

4. CITIZEN DIGITAL ARCHIVING

Libraries are still developing strategies for archiving all digital collections, personal or organizational, born-digital or converted from paper. What all collecting organizations know is that they cannot do it all themselves.

The U.S. National Archives has not only issued a public call for citizen archivists to help explore the collections[13], they hosted an event – "Are You In? Citizen Archivists, Crowdsourcing, and Open Government[14] -- to encourage participation in all forms of citizen projects. The open government movement in the United States has inspired people to participate in the digitization and transcription of documents to increase their discoverability and use. The best known individual in this realm is Carl Malamud, whose non-profit publicresource.org[15] is working toward the publication of public domain information from local, state, and federal government agencies. Among his best-known efforts are the digitization of over 550 government and publishing a 5 million page crawl of the Government Printing Office.

The New York Public Library is among the most recent organization to crowdsource a digital collection transcription effort, aiming to transcribe the dishes listed on 40,000 menus[16]. The earliest of these online efforts was likely the Fine Arts Museums of San Francisco Imagebase tagging project that began in 1997. One of the most extensive is the Flickr Commons, where more than fifty international organizations have shared images in order to crowdsource improvements to the metadata[17].

As vital as these efforts are for the improvement of digital collections, it is citizen efforts in the identification and acquisition of collections that we must encourage and collaborate with. We are all familiar with the efforts of the Internet Archive[18] which started out as a grassroots effort to archive the web, but now provides a repository for all manner of personal digital collections, from texts to video and audio to software and games.

We are also familiar with the work of Rick Prelinger, and his efforts to build a film archive and a library books, periodicals, printed ephemera and government documents[19]. His interest in creating a collection of ephemera, from non-theatrical films to zines to maps and plans, from textbooks to maps to government documents, coupled with his opening his library to the public, hosting traveling public showings from his film archive, digitizing portions of the collection and sharing it with the Internet Archive and the Library of Congress, has inspired likely thousands of others to build their own collections, both physical and digital.

Perhaps the least well-known, outside a group of rabid supporters, is Jason Scott, the founder of the related initiatives textfiles[20], ArchiveTeam[21], and URLTeam[22] among others. His cooperative, volunteer projects are an example of true grassroots digital archiving, stepping in to save marginalized collections, including bulletin boards and software manuals and web sites of imminent risk of shutdown. Scott is one of the most eloquent speakers on the risk of loss:

> "A wonderful thing happened in the 1980s: Life started to go online. And as the world continues this trend, everyone finding themselves drawn online should know what happened before, to see where it all really started to come together and to know what went on, before it's forgotten."[23]
>
> "Archive Team is a loose collective of rogue archivists, programmers, writers and loudmouths dedicated to saving our digital heritage. Since 2009 this variant force of nature has caught wind of shutdowns, shutoffs, mergers, and plain old deletions - and done our best to save the history before it's lost forever. Along the way, we've gotten attention, resistance, press and discussion, but most importantly, we've gotten the message out: *IT DOESN'T HAVE TO BE THIS WAY*."[24]

"Official" collecting institutions cannot afford to ignore the efforts of these citizen archivists, and, in fact, we must collaborate with them. A citizen digital archivist is often the most knowledgeable and motivated to save otherwise unrecognized at-risk materials. We need to encourage such efforts as well as recognizing them, incorporating them and their work into our distributed digital preservation community.

5. CONCLUSIONS

We need to recognize the importance of individual efforts in preservation. There is a clear role for the enthusiast in identifying what needs to be curated. For example, nineteenth-century collectors of books bound their collections for their use, and to share their collections with others which inadvertently contributed to future preservation. A more recent model is wikipedia, in which enthusiasts share their personal, curated knowledge. Another example is the work of amateur genealogists in sharing

[12] See http://www.digitalpreservation.gov/you/

[13] See http://blogs.archives.gov/online-public-access/?p=2661

[14] See http://blogs.archives.gov/aotus/?p=2938

[15] See https://public.resource.org/

[16] See http://menus.nypl.org/

[17] See http://www.flickr.com/commons?phpsessid=ea7b4da468f5935f24b65f41dbfc356f

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

[19] See http://prelingerlibrary.blogspot.com/.

[20] http://www.textfiles.com/

[21] http://www.archiveteam.org/index.php?title=Main_Page

[22] http://urlte.am/

[23] http://www.textfiles.com/statement.html

[24] http://www.archiveteam.org/index.php?title=Main_Page

their curated family histories. The personal role in preservation is significant--some people care enough to keep stuff alive; institutions may not, often because they are not aware of the importance of the content to one or more communities.
Is it necessary for an institution to accomplish all digital preservation? We do not need to do it all ourselves, and we already recognize that we cannot. In some cases, our institutions may be in the way. We need to create a new sense of sensitivity at our organizations to grassroots efforts. There are individuals curating at the fringes of our communities, but embedded within their communities. Our organizations need to give them support, and help them with a preservation strategy, whether it is guidance to individuals or collaboration with collecting institutions.

In the same way that institutions have encouraged citizen archiving of physical collections, we must encourage citizen digital archiving. First, we must educate ourselves about personal archiving requirements, recognizing the widespread need for skills in the community to work with a wide variety of formats and technologies needed to export and retrieve content from its silos – email programs, online images sites, blogs, and social media – and from potentially obsolete media. We must then undertake personal digital archiving education in our local communities. Hold personal digital archiving events that provide instruction on digitization and training on the best ways to preserve born digital and digitized files. Introduce people to the need to archive the web and to the use of web archiving tools and services. Partner with public libraries, historical societies, genealogical groups, and local history museums. We should reach out to the communities that are archiving their specialized histories, providing that same archiving advice and expertise, encouraging them to steward their collections, and seeking out and accepting those collections when they become available to us. We must also reach out to the vendors and software developers that create the tools used across multiple communities, encouraging them to use more preservable format standards and build export functionality that allow more portable, personal control over personal digital content. In this way, we are encouraging better personal stewardship and improved preservation of what will certainly become our future collections.

There is a clear mandate that libraries, archives and museums must assume responsibility for educating the public about strategies for personal digital archiving and personal curation, and for exploring new approaches to processing, preservation, and access. Personal digital items initially have value to the individuals who generated them, but once those items are transferred to a collecting institution they will have a collective value to society. We also have a less well-established mandate to work with those citizen archivists who are taking the initiative to save large collections of digital ephemera and the at-risk output of underserved communities. Large amounts or small, all digital materials have the potential to be vital to cultural history studies in the future.

6. REFERENCES

[1] Bowman, S. and Willis, C. 2003. *We Media: How Audiences are Shaping the Future of News and Information.* The Media Center at the American Press Institute, Reston, Virginia. http://www.hypergene.net/wemedia/weblog.php

[2] Cox, Richard J. 2009. Digital Curation and the Citizen Archivist. *Digital Curation: Practice, Promises & Prospects*. pp. 102-109.

[3] Gemmell, Jim, Gordon Bell and Roger Lueder. 2006. MyLifeBits: A Personal Database for Everything. *Communications of the ACM*, 49, 1 (January 2006), 88–95.

[4] Gordon Bell and Jim Gemmell. 2007. A Digital Life, *Scientific American*, 296, 3 (March 2007), 58-65. http://www.scienceandsociety.org/web/Library_files/A.Digital.Life.pdf

[5] Wilkinson, Alec. 2007. Remember This? A Project to Remember Everything We Do in Life. *The New Yorker*, (May 28, 2007). http://www.newyorker.com/reporting/2007/05/28/070528fa_fact_wilkinson

[6] Bush, Vannevar. 1945. As We May Think. *The Atlantic Monthly*, 176, 1 (July 1945), 101-108. http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/3881/

[7] Marshall, C.C., McCown, F., and Nelson, Michael L. 2007. Evaluating Personal Archiving Strategies for Internet-based Information. *Proceedings of Archiving 2007*. (Arlington, Virginia, May 21-24, 2007), Society for Imaging Science and Technology, 151-156. http://research.microsoft.com/apps/pubs/default.aspx?id=63770

[8] Jones, William. 2004. Finders, keepers? The present and future perfect in support of personal information management". *First Monday*, 9 3 (March 1, 2004) http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1123/1043

[9] Teevan, Jaime, William Jones, and Benjamin B. Bederson. 2006. Personal information management. *Communications of the ACM*, 49, 1 (January 2006), 40-43.

[10] Jones, W. 2007. Personal Information Management. *Annual Review of Information Science and Technology*, 41, 1 (2007), 453–504.

[11] Jones, W. 2007. *Keeping Found Things Found: The Study and Practice of Personal Information Management*. San Francisco, CA: Morgan Kaufmann.

[12] Jones, W., & Teevan, J. 2007. *Personal Information Management.* Seattle, WA: University of Washington Press.

[13] Marshall, C.C. Rethinking Personal Digital Archiving, Part 1: Four Challenges from the Field. *DLib Magazine*, 14, 3/4 (March/April 2008). http://www.dlib.org/dlib/march08/marshall/03marshall-pt1.html

[14] Marshall, C.C. Rethinking Personal Digital Archiving, Part 2: Implications for Services, Applications, and Institutions. *DLib Magazine*, 14, 3/4 (March/April 2008). http://www.dlib.org/dlib/march08/marshall/03marshall-pt2.html

[15] Beagrie, Neil. 2005. Plenty of Room at the Bottom? Personal Digital Libraries and Collections. *D-Lib Magazine*, 11, 6 (June 2005) http://www.dlib.org/dlib/june05/beagrie/06beagrie.html

[16] Williams, Pete, Katrina Dean, Ian Rowlands and Jeremy Leighton John. 2008. Digital Lives: Report of Interviews

with the Creators of Personal Digital Collections. *Ariadne* Issue 55 (30-April-2008). http://www.ariadne.ac.uk/issue55/williams-et-al/

[17] John, Jeremy Leighton. Adapting Existing Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools. Paper presented at iPRES 2008: The Fifth International Conference on Preservation of Digital Objects, London, UK, September 29-30, 2008. http://www.bl.uk/ipres2008/presentations_day1/09_John.pdf

[18] Paradigm project, *Workbook on Digital Private Papers*, 2005-7. http://www.paradigm.ac.uk/workbook/.

[19] Kirschenbaum, Matthew G., Erika Farr, Kari M. Kraus, Naomi L. Nelson, Catherine Stollar Peters, Gabriela Redwine, and Doug Reside. 2009. *Approaches to Managing and Collecting Born-Digital LiteraryMaterials for Scholarly Use*. White Paper. Washington, DC: National Endowment for the Humanities. http://www.neh.gov/ODH/Default.aspx?tabid=111&id=37

[20] John, Jeremy Leighton, Ian Rowlands, Peter Williams, and Katrina Dean. 2010. *Digital Lives: Personal Digital Archives for the 21st Century: An Initial Synthesis.* Digital Lives Research Project (3 March, 2010). http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf

[21] Matthew G. Kirschenbaum, Matthew G., Richard Ovenden, Gabriela Redwine, and Rachel Donahue. 2010. *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Washington D.C.: Council on Library and Information Resources.

[22] McDonough, Jerome P., et al. 2010. *Preserving Virtual Worlds: Final Report*. https://www.ideals.illinois.edu/handle/2142/17097

[23] Hafner, Katie. 2004. Even Digital Memories Can Fade. *New York Times* (November 10, 2004). http://www.nytimes.com/2004/11/10/technology/10archive.htm

[24] Hesseldahl, Arik. 2005. How To Preserve Photos For 500 Years. Forbes (April 14, 2005). http://www.forbes.com/2005/04/14/cx_ah_0414photo.html

[25] Cohen, Patricia. 2010. Fending Off Digital Decay, Bit by Bit. *The New York Times*. (March 15, 2010) http://www.nytimes.com/2010/03/16/books/16archive.html

[26] Fox, Stuart. 2010. Digital Age Presents New Problems for Historians. *Tech News Daily* (July 16, 2010). http://www.technewsdaily.com/information-overload-digital-age-presents-new-problems-for-historians-0867/

[27] Paul-Choudhury, Sumit. 2011. Digital legacy: Respecting the digital dead. *New Scientist* (May 6, 2011). http://www.newscientist.com/article/dn20445-digital-legacy-respecting-the-digital-dead.html

[28] Carroll, Evan and John Romano. 2010. *Your Digital Afterlife: When Facebook, Flickr and Twitter Are Your Estate, What's Your Legacy*. Berkeley, CA: New Riders Press. http://www.yourdigitalafterlife.com/

[29] Walker, Rob. 2011. Cyberspace When You're Dead. *New York Times Magazine* (January 9, 2011). http://www.nytimes.com/2011/01/09/magazine/09Immortality-t.html

[30] Bell, Gordon. 2000. Dear Appy, how committed are you? Signed, lost and forgotten data. *Ubiquity* (February 1 - February 28, 2000). http://ubiquity.acm.org/article.cfm?id=334401

[31] Wright, Gary T. 2010. Preserving Your Family History Records Digitally. *FamilySearch*. (October 2010). https://wiki.familysearch.org/en/White_Paper:_Preserving_Your_Family_History_Records_Digitally

The Enduring Value of Social Science Research:
The Use and Reuse of Primary Research Data

Amy Pienta
ICPSR, University of Michigan
PO Box 1248
Ann Arbor, MI, USA 48106-1248
011.1.734.615.7957
apienta@umich.edu

George Alter
ICPSR, University of Michigan
PO Box 1248
Ann Arbor, MI, USA 48106-1248
011.1.734.615.7652
altergc@umich.edu

Jared Lyle
ICPSR, University of Michigan
PO Box 1248
Ann Arbor, MI, USA 48106-1248
011.1.734.763.6075
lyle@umich.edu

ABSTRACT
The goal of this paper is to examine the extent to which social science research data are shared and assess whether sharing is associated positively with number of publications resulting from the research data. We construct a database from administrative records containing information about thousands of social science studies that have been conducted over the last 40 years. Included in the database are descriptions of social science data collections funded by the National Science Foundation and the National Institutes of Health. Using a subset of these awards, we conduct a survey of principal investigators (n=1,021). We find that very few social science data collections are preserved and disseminated by an archive or institutional repository. Informal sharing of data in the social sciences is much more common. The main analysis examines publication metrics that can be tied to the research data collected with NSF and NIH funding – total publications, primary publications (including PI), and secondary publications (non-research team). Multivariate models of the count of publications suggest that data sharing, especially sharing data through an archive, is associated with many more times the publications compared to not sharing data. This finding is robust even when the models are adjusted for PI characteristics, grant award features, and institutional characteristics.

Categories and Subject Descriptors
Scientific databases, Statistical databases, Economics, Sociology

General Terms
Management, Measurement, Economics

Keywords
Research Data Sharing, Scientific Productivity, Digital Preservation

1. INTRODUCTION
Federal funding for scientific research has always been a highly competitive endeavor with only a small proportion of research grant submissions receiving awards from the National Institutes of Health (NIH) each year. The impact of a funded research project is measured, partly, by the research productivity of the PI and his or her research team who publish findings from primary data collection activities. Increasingly, NIH and the National Science Foundation (NSF) have become interested in data sharing as a means of supporting the scientific process and ensuring the highest return on competitive investments. However, there has been little investigation of research productivity that extends beyond the primary analysis of hypotheses outlined in the original data collection project. We proposed to redress this gap by examining data-related research productivity of the research team and secondary use by others.

This research question is particularly salient for the social sciences because social science disciplines have been among the earliest to organize efforts to share research data. Avenues for sharing data have been fairly well known, especially in the social science disciplines of political science, sociology and economics. Social science research occurs in other social and behavioral disciplines, as well. So, there is tremendous heterogeneity in data sharing in the social sciences.

The largest share of social science research is conducted with federal support. The National Science Foundation (NSF) and the National Institutes of Health (NIH) have supported a significant share of social science data collections and the trend continues today (Alpert, 1955; Alpert 1960; Kalberer, 1992). This paper focuses on analyzing information from grant awards made by NSF and NIH making it possible to enumerate the bulk of the major social science data collections that exist today. Also, NSF and NIH keep electronic records about grant awardees that have been culled into a single database useful for understanding the scope and breadth of social science research that has produced research data. Thus, this research topic is both timely and practical.

2. BACKGROUND
Data sharing has been an important topic of debate in the social sciences for more than twenty years, initially spurred by a series of National Research Council Reports and more recently the publication of the National Institutes of Health Statement on Sharing Research Data in February 2003 (NIH 2003). Despite this formal written statement from NIH and a similar one from the National Science Foundation (NSF-SBE n.d.) that give official support for the long held expectations placed on grantees to share their research data, little is known about the extent to which data collected with support from NIH or NSF have been shared with other researchers. The limited work done suggests considerable variability in the extent to which researchers' share and archive

iPRES2011, Nov. 1–4, 2011, Singapore.

research data. Our research fills this gap in knowledge and creates a research database for answering these questions.

NIH's policy is designed to encourage data sharing with the goal of advancing science. The benefits of sharing data have been widely discussed and understood by researchers for years. An important part of Kuhn's (1970) scientific paradigm is the replication and confirmation of results. Sharing data is at the core of direct replication (Anderson et al. 2005; Kuhn 1970; Freese 2006). The foundation of the scientific process is that research should build on previous work, where applicable, and data sharing makes this possible (Bailar 2003; Louis, Jones & Campbell 2002). The argument has been made, and there is some evidence to support it, that sharing data and allowing for replication makes one's work more likely to be taken seriously and cited more frequently (King et al., 1995). In fact, Glenditsch, Petter, Metelits, and Strand (2003: 92) find that authors who make data from their articles available are cited twice as frequently as articles with "no data but otherwise equivalent credentials, including degree of formalization."

Additionally, the nature of large datasets virtually guarantees that a single researcher or group of researchers will not be able to use the dataset to its full potential for a single project. It may be the case that those who collect the data are not the best at analyzing them beyond basic descriptive analyses (Bailar 2003). Sharing data in this way ensures that resources spent on data collection are put to the best use possible and the public benefit is enhanced.

Finally, the use of secondary data is crucial in the education of undergraduate and graduate students (Fienberg, 1994; King, 2006). It is not feasible for students in a semester-long course to collect and analyze data on a large scale. Using archived datasets allows students to gain experience firsthand. Instructors can use the metadata accompanying shared data to teach students about "good science" and the results obtained from even simple analyses to illustrate the use of evidence (data) in support of arguments (Sobal 1981).

2.1 Policies about data sharing

Most institutes and organizations that finance research, especially data collection, have a policy about sharing data once the initial project is completed. The National Institutes of Health (NIH 2003) and National Science Foundation (NSF-SBE n.d.), for example, require a clearly detailed plan about data sharing as part of research proposals submitted for review. Plans must cover how and where materials will be stored; how access will be given to other researchers; and any precautions that will be taken to protect confidentiality when the data are made public. These requirements are not, however, evaluated in the review process nor are there formal penalties for non-compliance after the award. Most professional organizations also include a statement in their "best practice" or ethics guidelines recommending that research reports be detailed enough to allow for replication, and that data and assistance be made available for replication attempts (e.g., American Sociological Association, American Psychological Association, American Association for Public Opinion Research).

In addition to such general statements that data collected with public funds must be shared with other researchers and that individuals should be willing to assist others replicating their work, some fields, such as Economics, have taken steps to make the data sharing policy more concrete. In an attempt to allow for direct replications as well as full-study replications, the American Economic Review and other major economics journals have instituted the practice that any article to be published must be accompanied by the data, programs used to run the analyses, and clear, sufficient details about the procedures prior to publication (Freese 2006; Anderson et al. 2005). The requirement to include not only the data but also statistical code written to perform analyses requires that individual researchers thoroughly and carefully document decisions made during the analysis stages of the project and allows other researchers to more easily use these as starting points for their own work. This has led to increased use and citation of work that has been published in journals where this type of information is required (Anderson et al. 2005; Glenditsch et al. 2003).

2.2 Sharing Social Science Data

Data are currently shared in many different ways ranging from formal archives to informal self-dissemination. Data are often stored and disseminated through established data archives. These data generally reach a larger part of the scientific community. Also, data in formal archives typically include information (metadata) about the data collection process as well as any missing data imputations, weighting, and other data enhancements. These archiving institutions have written policies and explicit practices to ensure long-term access to the digital assets that they hold, including off-site replication copies and a commitment to the migration of data storage formats. These are the characteristics that define data archives.

Another tier of data archives have more narrowly focused collections around a particular substantive theme such as the Association of Religion Data Archives (www.thearda.com). The data in these kinds of thematic archives are not necessarily unique, though some of their holdings are, but the overlap between archives makes data available to broader audiences than might be captured by a single archive. The ARDA, for instance, has a broader non-scientific audience who are interested in analysis and reports as well as the micro-data files for reanalysis. These archives expend resources on the usability of the collection and make some commitment to long-term access through migration and back-ups.

Some data archives are designed solely to support the scientific notion of replication. Journal-based systems of sharing data have become popular in Economics and other fields as a way of encouraging replication of results (Anderson et al. 2005; Glenditsch et al. 2003). The longevity of these collections is sometimes more tenuous than the formal archives particularly if the sustainability of their archival model relies on a single funding source.

Some examples of less formal approaches include authors who acknowledge they will make their data available upon request or who distribute information or data through a website. Researchers often keep these sites up to date with information about findings from the study and publication lists, in addition to data files and metadata. These sites are limited to those who know about the study by name or for whom the website has shown up in a Web search (see also Berns, Bond & Manning 1996). Typically, the commitment to preserving this content lasts only as long as the individual has resources available.

2.3 The Reluctance of Researchers to Archive Data

The time and effort required to produce data products that are useable by others in the scientific community is substantial. This extra effort is seen by many as a barrier to sharing data (Birnholz & Bietz 2003; Stanley & Stanley 1988). In addition to the actual data, information must be added to assist secondary users in identifying whether the data would be of value to them and in the

analysis and interpretation of results. Such metadata includes complete descriptions of all stages of the data collection process (sampling, mode of data collection, refusal conversion techniques, etc.) as well as details about survey question wording, skip patterns and universe statements, and post-data processing. All of these factors allow subsequent researchers to judge the quality of the data they are receiving and whether it is adequate for their research agenda. Therefore, substantial effort is required of those sharing data, while the lion's share of the benefits seem to accrue to the secondary user.

Another significant barrier in the sharing of data is the risk of breaching the confidentiality of respondents and the potential for the identification of respondents (Bailar 2003). The issue of protecting confidentiality has become more salient as studies collect information about social context, which may include census tract or block group identification to allow researchers to link the data collected with information about the context. Not only are data about social and community contexts being collected and included in datasets but also global positioning coordinates and information about multiple members of a household, all of which could make identification of any single individual easier. Additional information about biomarkers and longitudinal follow up are also hallmarks of new data collection efforts. Both methodological innovations make it more difficult for Institutional Review Boards to allow for the wide redistribution of data.

Other reasons individuals give for withholding data include wanting to protect their or their students' ability to publish from the data as well as the extra effort involved in preparing data for sharing (Louis et al. 2002). Retaining the ability to publish from one's data is a significant concern among scientists, both for fear of others "scooping" the story and that others will find mistakes in their attempt to replicate results (Anderson et al. 2005; Bailar 2003; Freese 2006; Bachrach & King 2004).

Current publication and academic promotion practices act as another barrier to sharing data – or, put another way, those who "hoard" their data are likely to be rewarded more than those who "share". There are often few, if any, rewards to sharing data, especially given the expense in terms of time and effort required to prepare clean, detailed data and metadata files. Researchers are not typically rewarded for such behavior, particularly if the time spent on data sharing tasks infringes on one's ability to prepare additional manuscripts for publication. Academic culture does not support the scientific norm of replication and sharing with tangible rewards. (Anderson et al. 2005; Berns et al. 1996). As an example, in discussing the notion that researchers might share not only data but also analytic/statistical code, Freese (2006:11) notes that a typical reaction to a "more social replication policy would be to expend less effort writing code, articulating a surprisingly adamant aversion to having [one's] work contribute to others' research unless accompanied by clear and complete assurance in advance that they would be credited copiously for any such contribution." It is unlikely that attitudes about data sharing will change without strong leadership and examples set by senior scientists and the commitment of scientific institutions such as universities and professional societies who facilitate and enforce such sharing (Berns et al. 1996).

2.4 Extending Research Productivity to Include Data Reuse

Research productivity is often thought of as something that scientists accomplish by publishing their research discoveries. The second part of research productivity is not how many times your ideas are published, but also how often the idea is cited in the work of others (Matson, Gouvier, Manikam 1989). This is an analysis of citation counts of a scientist's publications – how widely cited their publications are. Thus, the impact of a scientist's scholarship is derived directly from their own published work. However, there has been movement in the scientific academy to recognize the importance and value of research data. We consider the possibility that research data may have enduring value on scientific progress as scientists use and reuse research data to draw new analysis and conclusions. This idea is rooted in the idea of a data life cycle – where research data can often have use beyond its original designed purpose (Jacobs and Humphrey 2004). This is not farfetched given that research productivity measures have also been used to assess institutional productivity across universities (Toutkoushian, Porter, Danielson, and Hollis, 2003). Here, we consider the research productivity resulting from research data collected by a scientist with federal funding.

In summary, while the social sciences share in the normative expectation that research data must be shared to foster replication and reanalysis, there is little to suggest that it is a wide spread practice. Federal institutions and professional organizations underscore these normative expectations with implicit and explicit sharing policies. The advantages of sharing data with the research community are large and cumulative. Yet, with the exception of leading journals in Economics, there are few cases in which these normative statements are coupled with penalties or incentives to reinforce them. The institutional, financial, and career barriers to data sharing are substantial as noted. What remains an open empirical question is the extent of data sharing across social science disciplines and the value this has for the social sciences.

3. Methods

To address this question we construct a database of research projects -- the 'LEADS' database -- is comprised of social and behavioral science awards made by NSF and NIH. From the National Science Foundation online grants database, we include in our study research grant awards that matched prominent search terms relating to the social sciences (We used the following search terms to select possible awards from NSF for inclusion in LEADS: SOC*, POLIT*, and/or STAT*). We further restrict this set of awards to awards that include descriptions of research activity that (1) relate to the social and /or behavioral sciences and (2) reflect original (or primary) data collection (including assembly of a new database from existing or archival sources). From the National Institutes of Health online CRISP (Computer Retrieval of Information on Scientific Projects) database (http://crisp.cit.nih.gov/), we include extramural research grant awards from the top 10 NIH institutes engaging in social and behavioral research. In additional to screening for social and behavior science content in these awards, these awards also were restricted to the collection of original quantitative data. This strategy differs from the NSF award review in that strictly qualitative studies were not identified as such and excluded from LEADS (because the database was constructed from an earlier project that explicitly excluded qualitative studies). Because mixed method studies were screened in - the potential impact of this difference is small.

Of the 235,953 eligible NSF and NIH awards in the LEADS database, 12,464 matched our initial screening criteria (i.e., social/behavior science & collected research data). We then select awards from 1985-2001 (n=7,040). We selected this range of years because we wanted to inquire about completed research that could have led to publications and data archiving. But, we did not want to select awards that were completed so long ago that recall of information about the publications related to the award would

be unreasonable. From this set of awards, we found 4,883 unique principal investigators (PIs). We attempted to invite all 4,883 PIs to complete a web survey (excluding deceased PIs and PIs where we could not verify an email address).

The PI survey consisted of questions about research data collected, various methods for sharing research data, attitudes about data sharing and demographic information. PIs were also asked about publications tied to the research project including information about their own publications, research team publications, and publications outside the research team. We received 1,217 responses (24.9% response rate). For the analytic sample we select PIs and information about their research award if (1) they confirm they collected research data as part of the selected award (86.6% of the responses) and (2) they did not collect data for a dissertation award.

3.1.1 Publication Measures

Research productivity is typically assessed by either citation or publication analysis. The outcome measures used in this analysis are various measures of publication counts. Publication counts are based on self-reported information provided by PIs of the research grant awards at NSF and NIH. PIs are asked to report number of publications related to the data they collected, including estimates for: own publications, publications of the research team, extant publications not related to the research team, and the number of publications (in each of the three previous categories) that include students. We include in this analysis count of publications where the PI is one of the authors (range 0 – 100). This is one measure of primary publications. A second measure of primary publications is created that also includes counts of publications where the PI may or may not be an author, but at least one member of the research team is an author (range 0-350). Secondary publications are publications where none of the original research team (PI, co-investigators, students or other researchers) is an author or co-author of the publication (range 0-700). This measure indicates the extent of reuse (or secondary use) of research data beyond its original collection purpose. Next, total publication count is constructed by adding count of all primary publications with count of secondary publications (range=0-713). Finally, the number of publications where a student was author or co-author is defined (range 0-160).

Because the publication measures are self-report measures, we conducted a separate publication search (using Web of Science, Google) for a sub-set of awards to verify that PI self-reports were correlated with an objective set of publication counts. In analyses not reported here, we find that self-report and objective publication counts are highly correlated. On average, PIs report more publications regardless of the publication count measure (primary, secondary and so on). Thus, they tend to over report all kinds of publications, not just their own

3.1.2. Data Sharing Status

The main independent variable used in the analysis is self-reported data sharing status. We ask the question about data sharing in the PI survey. PIs are asked if they have ever shared data from their selected award through either an archive or more informal venue. Informal data sharing is a summary of information reported by the PI indicating either data were made available at the request of another researcher and/or they distributed the data through a personal or departmental website. Data sharing status is defined as whether research data have been shared (1) formally through a data archive (or institutional repository), (2) informally, not through a data archive (including shared upon request, personal website, departmental website), or (3) not shared.

3.1.3 PI Characteristics

To ensure that data sharing is not "standing in" for other known predictors of productivity, we include covariates describing characteristics of the individuals who collect the data, the award mechanism used to fund the data, and the institutional home of the original data collection. Research productivity has been linked to departmental prestige (Long 1978), age (Levan and Stephan 1991) and gender (Penas and Willet 2006) among other factors. We begin by describing PI characteristics we are able to measure.

We expect that characteristics of the PIs themselves will be associated with both data sharing status and various publication counts. Some researchers have more time for archiving and publishing whereas others may be more likely to engage in training and service. We attempt to control for this by including various social and demographic characteristics of the PIs in the models. The gender of the PI is male (=1) or female (=0). The self-reported race/ethnicity of the PI is defined as white (=1) versus non-white (=0). Age (in years) at time of initial award is calculated by subtracting year of birth from year at start of initial award (range 27-75). Self-reported faculty status/rank at time of initial award is defined as senior (tenured faculty), junior (untenured faculty), and non-faculty (including students, postdocs, research staff). Self-reported discipline is classified from an open ended question and collapsed into the following categories: (1) health sciences (nursing, medicine, public health) and psychology, (2) core social science (political science, sociology, and economics), and (3) other social science-related discipline (anthropology, film, communications). Finally, the number of federal grants awarded throughout one's career is defined as number of self-reported federal research grants (range 1-100).

3.1.4 Institutional Characteristics

Next, we construct a set of measures about the institutions awarded the research grant – the institution of the PI at time of initial award. First, we use the Carnegie Classification to differentiate research institutions from non-research institutions. Research institutions include research universities, doctoral granting universities, and medical schools/centers. Non-research institutions include 2- and 4- year colleges, colleges and universities granting Master's degrees, professional institutions and tribal colleges. Other institutions not classified under Carnegie are divided into private research organizations and other non-Carnegie institutions. A second institutional characteristic defined is the region where the institution is located (northeast, south, midwest and west).

3.1.5 Grant Award Characteristics

First, we differentiate awards made by the National Science Foundation (=0) from the National Institutes of Health. NSF has had in place a data sharing policy for a longer time and it is expected that data will be shared and archived more frequently when funded by NSF. The other award measure is the duration of the award, measured in years (range =0-8 years).

3.2. Analysis Plan

Descriptive statistics are calculated using univariate and bivariate statistics. Because the outcome measures are publication counts, Poisson regression models are estimated. Overdispersion led us to the choice to estimate negative binomial regression models of publication counts for longitudinal data (offset by the amount of time between initial award and the survey). We estimate two sets of models for each outcome. First, we estimate models that

include only a three category data sharing status measure. The second set of models adds the various PI, institution, and award characteristics. We do not include any covariates in the final models shown (model 2) that were not statistically significant across all outcomes. The hierarchical set of models (model 1 and model 2) allows us to understand the extent to which differences by data sharing status might be attributed to other characteristics of PIs, institutions and the awards.

4. RESULTS

Descriptive sample characteristics are presented in Table 1. The sample of PIs is fairly evenly divided between males (51.9 %) and females (48.1%). The majority of the sample is white (86.8 %) and tenured (54.3 %). Only 20 percent of the PIs is non-faculty. The mean number of Federal grants the PIs have been awarded throughout their careers is 6.2. The majority of PIs come from either the psychological or health sciences (62.5%). Just over a quarter of the sample are PIs in the core social science disciplines (sociology, economics and political science).

Table 1. Descriptive Sample Characteristics (n=930)

	Total	Range
PI Characteristics		
Female (%)	48.1	
White (%)	86.8	
Age @ award time(mean)	43.4	27-75
Faculty Status @ Award - Senior (%)	54.3	
Faculty Status @ Award - Junior (%)	25.7	
Faculty Status @ Award) - Non-Fac (%)	20.0	
Discpline - Core Social Science	25.5	
Discipline - Psychology & Health	62.5	
Disciple - Other	12.0	
# Fed Grants in Lifetime (mean)	6.2	1-100
Institutional Characteristics		
Region - NorthEast (%)	36.0	
Region - MidWest (%)	23.7	
Region - South (%)	21.6	
Region - West (%)	18.7	
Carnegie-Research (%)	78.7	
Carnegie-Non Research (%)	6.5	
Carnegie-Other, PRO (%)	12.4	
Carnegie-Other, Other (%)	2.5	
Grant Characteristics		
NSF Award (%)	27.3	
Duration of Initial Award, Years	3.1	0-8
Data Sharing Status		
Shared Formally, Archived	11.5	
Shared Informally, Not Archived	44.6	
Not Shared	43.9	

PIs are represented in all four major regions of the U.S. The largest numbers of grant awards are made to institutions located in the northeast (36%) and the fewest number of grant awards are made to institutions located in the west (18.7%). The vast majority of PIs of the research grant awards are working at institutions classified by the Carnegie classification as research institutions (78.8%). The second largest institution type represented in the PI survey is private research organizations (12.3%). Few awards were made to non-research institutions and other types of organizations not classified by Carnegie (6.5% and 2.5% respectively). Only 27.3 percent of the awards come from the National Science Foundation with majority coming from the National Institutes of Health (72.7%). The mean duration of an award is 3.1 years. Few awards produce research data that are shared formally – either in a data archive or institutional repository (11.5%). Of the rest, half the data from the awards are shared informally, not in an archive (44.6%) and half are not being shared beyond the research team (43.9%).

Table 2. Bivariate Relationships: Data sharing status by PI Characteristics, Institutional Characteristics, and Grant Award Characteristics

	Shared Formally, Archived (n=111)	Shared Informally, Not Archived (n=415)	Not Shared (n=409)	p-value
PI Characteristics				
Female (%)	15.1	42.2	42.7	***
Male	7.6	47.2	45.2	
White (%)	12.0	45.5	42.5	*
Nonwhite	8.1	39.0	52.9	
Age @ Award (mean)	44.3	43.4	43.1	
Fac Stat@Award-Senior (%)	14.7	45.7	39.6	***
Fac Stat@Award-Junior (%)	7.1	48.5	44.4	
Fac Stat@Awrd-NonFac (%)	8.6	36.6	54.8	
Discipline - Core Social Sci	27.0	48.5	24.5	***
Discipline - Psych & Health	4.7	42.9	52.5	
Disciple – Other	14.3	45.5	40.2	
# Federal Grants (mean)	7.3	6.3	5.8	
Institutional Characteristics				
Region - Northeast (%)	29.7	39.3	34.5	*
Region - Midwest (%)	15.5	43.6	40.9	
Region - South (%)	8.5	45.3	46.3	
Region - West (%)	13.8	37.4	48.9	
Carnegie-Research (%)	11.9	44.7	43.4	
Carnegie-Non Research (%)	8.3	38.3	53.3	
Carnegie-Other, PRO (%)	12.2	47.0	40.9	
Carnegie-Other, Other (%)	4.4	47.8	47.8	
Grant Award Characteristics				
NSF Award (%)	22.4	43.7	33.9	***
NIH Award	7.4	45.0	47.6	
Duration of Award, Years	2.9	3.3	2.9	

* p<.1; ** p<.05; ***p<.01 (p-values for chi square tests)

4.1 Characteristics of PIs Sharing Research Data.

Turning to Table 2, we next examine how various characteristics of the PIs, institutions and grant awards are related to data sharing status. Women are more likely to archive data than men (12.0% and 8.1% respectively; chi-square is statistically significant). We see that senior faculty are more likely than others to archive data (12.0%) – nearly twice as often as junior faculty (7.1%) and non-faculty (8.6%). There are strong disciplinary differences as well. The core social science disciplines archive data at the highest rate (27%). Psychologists and health scientist archive data least often (4.6%). PIs at institutions located in the south are also less likely to archive data (8.5%). However, the Carnegie classification of the institution awarded a research grant to collect data is not associated with data sharing status. Data funded by NSF research grant awards are nearly three times more likely to be archived than data funded by NIH.

Table 3. Bivariate Results: Data Sharing Status by Publication Counts

	Total n=935 Median	Archived n=111 Median	Informal n=415 Median	Not Shared n=409 Median
Primary Publications (w/ PI)	4	6	6	3
Primary Publications (w/any Research Team Member)	5	8	6	3
Secondary Publications (no Team Member)	0	0	0	0
Total Publications	5	10	7	4
Total Publications including Students	2	4	3	1

4.2 Data Sharing is Positively Associated with Number of Publications

Table 3 shows the distribution of various publication counts for the full sample and by data sharing status. The median number of publications for an award producing data that PIs author or co-author is 4. However, the median number of publications that PIs who archive data formally write is 6 – compared to PIs who do not share data (3 primary publications). Research teams are also more productive when they archive the data. The median number of research team publications is 8 when data are archived compared to 3 when data are not shared outside the research team.

Large numbers of research data produce no secondary publications beyond the PI and research team. Thus, across all categories, the median number of secondary publications is 0. For this outcome we examine the mean. A research grant award produces 2 secondary (non-research team) publications on average. However, when data are archived, 4 secondary publications are reported on average. We turn to the total number of publications next. A research grant award produces a median of 5 total publications. However, when data are archived a research grant award leads to a median of 10 publications. When data are shared informally a research grant is linked to a median of 7 publications. And, when data are not shared outside the research team, the research data lead to a median of only 4 publications overall. The same pattern is found for publications with student authorship as well.

Multivariate results are presented in Table 4. Dispersion differs from 0 across all outcomes and models leading us to estimate negative binomial regression models. Log-likelihood estimates are presented in Tables 4 & 5 (standard errors appear in parentheses). Both archiving data and sharing data informally are related positively to count of total publications (b=1.094 and b=1.020 respectively). Both associations are statistically significant (p<.01). This can be interpreted (by taking the exponential of the log-odds) as archiving data leading to 2.98 times more publications than not sharing data. When data are shared informally (compared to not shared at all), 2.77 times the number of publications are produced.

Turning to model 2, additional covariates are added to the model to account for potential differences in PIs, institutions, and the grant awards. The coefficients for archiving data and informally sharing data are positively associated with number of total publications in comparison to not sharing data at all. These two coefficients are smaller than in model 1, but still statistically significant. This can be interpreted (by taking the exponential of the log-odds) as archiving data leads to 2.42 times more publications than not sharing data. Informally sharing data leads to 2.31 times more publications than not sharing data at all. Thus, the effect of data sharing, formally or informally, is not explained by differences in the PI themselves, the awards or the institutions that were given the awards to conduct the research. Research productivity benefits clearly from data sharing, particularly archiving data.

Table 4. Multivariate Results: Negative Binomial Regression Models of Publication Counts

	Total # Publications, Self-Reported						Total # Secondary Publications, Self-Reported					
	Model 1			Model 2			Model 1			Model 2		
Data Sharing Status												
Primary Publications (w/ PI)	1.094	(0.123)	***	0.884	(0.128)	***	2.515	(0.415)	***	1.919	(0.443)	***
Shared Informally-Not Archived	1.020	(0.080)	***	0.837	(0.079)	***	2.375	(0.276)	***	1.565	(0.284)	***
Not Shared	ref			ref			ref			ref		
PI Characteristics												
Age at award				0.025	(0.004)	***				0.037	(0.016)	**
Discipline - Health and Psychology				-0.254	(0.102)	**				-0.977	(0.370)	***
Discipline - Other (vs Core Soc Sci)				-0.190	(0.130)					-1.107	(0.467)	***
Institutional Characteristics												
Carnegie-Non Res University				-0.685	(0.157)	***				-0.623	(0.584)	
Carnegie-Other				1.169	(0.246)	***				1.602	(0.840)	*
Carnegie-PRO (vs Res Univ)				0.230	(0.113)					1.230	(0.387)	***
Grant Award Characteristics												
NIH (vs. NSF)				0.075	(0.093)					-0.202	(0.358)	
Duration of Award, Years				0.163	(0.027)	***				0.115	(0.102)	
Intercept	1.646	(0.058)		0.199	(0.222)		-1.314	(0.206)	***	-2.418	(0.794)	***
Dispersion	1.186			1.052			13.649			11.241		

* p<.1; ** p<.05; ***p<.01

Other coefficients in the model demonstrate that being older at the time of award is associated with increasing log-odds of total publications. Being older at time of award may translate into a measure of writing and publishing experience – and in turn older PIs may have a publishing advantage that is not explaining by other factors. One of the surprising results is that faculty status (senior, junior, and non-faculty) at time of award was not statistically significant in the model. This covariate (and gender, race and number of federal grants) are not included in model 2.

Turning to model 2, additional covariates are added to the first model to account for potential differences in PIs, institutions, and the grant awards. The coefficient for archiving data is positively associated with secondary publication count in comparison to not sharing data at all, but is smaller than in model 1 (b=1.919 in model 2 compared to b=2.515 in model 1). This can be interpreted (by taking the exponential of the log-odds) as archiving data leads to 6.81 times more secondary publications than not sharing data. Both archiving and informal sharing are

Table 5. Multivariate Results: Negative Binomial Regression Models of Publication Counts

	Total # Primary Publications, Self-Reported						Total # Student Publications, Self-Reported					
	Model 1			Model 2			Model 1			Model 2		
Data Sharing Status												
Primary Publications (w/ PI)	0.620	(0.111)	***	0.729	(0.112)	***	0.700	(0.156)	***	0.936	(0.165)	***
Shared Informally-Not Archived	0.743	(0.073)	***	0.67	(0.069)	***	0.763	(0.103)	***	0.665	(0.100)	***
Not Shared	ref			ref			ref			ref		
PI Characteristics												
Age at award				0.024	(0.004)	***				0.022	(0.006)	**
Discipline - Health and Psychology				0.161	(0.089)	*				0.324	(0.125)	***
Discipline - Other (vs. Core Social Sci)				0.141	(0.113)					0.276	(0.165)	*
Institutional Characteristics												
Carnegie-Non Res University				-0.558	(0.142)	***				-0.888	(0.205)	***
Carnegie-Other				0.901	(0.211)	***				1.091	(0.335)	***
Carnegie-PRO (vs. Res Univ)				-0.216	(0.099)	**				-0.981	(0.147)	***
Grant Award Characteristics												
NIH (vs. NSF)				0.226	(0.081)	***				0.234	(0.113)	**
Duration of Award, Years				0.200	(0.024)	***				0.207	(0.034)	***
Intercept	1.444	(0.053)		-0.521	(0.206)		0.989	(0.075)	***	-0.982	(0.301)	***
Dispersion	0.902			0.75			1.832			1.559		

* p<.1; ** p<.05; ***p<.01

The other PI characteristic that affects total publications is PI's discipline. Compared to data collected by core social scientists, data collected by health scientists and psychologists have lower log-odds of leading to overall publications (b=-.254).

Only one measure of the institutional climate surrounding the award that produced data is retained in model 2. Carnegie classification is associated with total publication count. Compared to data collected at research universities, data collected at non-research institutions reduce the log-odds of overall publications (b=-.685). Data collected at other non-Carnegie classified institutions (but not private research organizations which were classified separately), compared to data collected at Carnegie research universities, are actually associated with increased log-odds of publications (b=.230). Finally, the greater the length of the initial award period the greater the log-odds of publication (b=.199).

The next set of models examines the number of secondary publications. Secondary publications are publications by researchers outside the research team. We find that secondary publications are also related to data sharing status. Both archiving data and sharing data informally are positively related (increase the log-odds) of secondary publications (b=2.515 and b=2.375 respectively). Both associations are statistically significant (p<.01). This can be interpreted (by taking the exponential of the log-odds) as archiving data leads to 12.37 times more publications than not sharing data. When data are shared informally (compared to not shared at all), 10.75 times the number of publications are produced.

odds of primary PI publications (b=.743). Both associations are statistically significant (p<.01). This can be interpreted (by taking the exponential of the log-odds) as archiving data leading to nearly 2 times more publications than not sharing data. Adding the additional covariates in model 2 does not explain the data sharing effects. In the last set of models we saw private research organizations (PROs) produce data that lead to greater numbers of secondary publications. Here, in model 2, we see that PROs produce data that lead to lower log-odds of primary PI publications compared to research universities (b=-.216). Also in this model, we see that NIH data increase the log-odds of primary publications compared to NSF data.

Much like the other publication metrics, the number of publications including students is related to data sharing status. Archiving (b=.700) and sharing data informally (b=.763) increase the log-odds of publications including students in comparison to not sharing data. Adding the additional covariates in model 2 does not explain data sharing differences.

5. CONCLUSIONS

The research database we constructed contains valuable information about a wide range of social science research data collected with support from the National Science Foundation and the National Institutes of Health. NSF and NIH awards typically lead to some of the largest investigator-initiated research activities in the U.S. and both institutions have had longstanding expectations that data collected with public money ought to be made available to the public and/or research community. In the social science research community, more so than in other basic disciplines, there have been longstanding avenues for archiving

and sharing data. Even with this advantage, we confirm that the majority of social science data are not archived publicly (88.5%). Informal data sharing, though much more common (44.6%), does not ensure that the scientific information collected with public funding has enduring value beyond its original primary publications.

One of the central questions stemming from this disparity is whether research productivity varies by data sharing. We find strong and consistent evidence that data sharing, both formal and informal, increases research productivity across a wide range of publication metrics. Data archiving, in particular, yields the greatest returns on investment with research productivity (number of publications) being greater when data are archived. Not sharing data, either formally or informally, limits severely the number of publications tied to research data. We hypothesize that some of the data sharing advantage would be explained by PI characteristics and characteristics of the institutions and grant awards. We find that although this is true, large persistent advantages in research productivity accrue when data are shared. Finally, we also include a large number of publication metrics to better understand how data sharing affects primary versus secondary publications. Data sharing is related to all publication metrics, even primary PI publications. However, data sharing has the largest effects on secondary publications, as expected. Data archiving, and informal data sharing, generate many more secondary publications than PI and research team exclusive use.

5.1 Limitations

The measures of publication counts in the paper are self-reported. This could lead to incorrect estimates of publications counts, particularly of secondary publications. However, the results reported here are consistent across counts of primary and secondary publications. Also, we collected publication counts based on our own citation search for a select number of grant awards. We confirm higher publication counts for data that are found to be archived (results available upon request from authors) with a more limited set of covariates.

Also, it is unclear whether larger numbers of primary publications lead to data sharing or if sharing data leads to more primary publications. While both are plausible, it is likely that the association we observe between data archiving and primary publications reflects the fact that PIs archive data when their research is complete and all primary findings are published. That said, we carefully selected a range of grant awards that would have been completed years ago.

Larger research projects probably lead to more publications and greater likelihood of data sharing. While we have included a measure of grant award duration to get at some of the variability in grant award size, a better measure of the size of the research project is total amount in dollars of the award. The largest social science data collections simply cost more money to collect, are intended for public dissemination, and have more information that would appeal to a larger number of scientists. Unfortunately this information is not available for NIH awards.

6. ACKNOWLEDGMENTS

We would like to acknowledge the National Library of Medicine (R01 LM009765). We also thank Darrell Donakowski, Myron Gutmann, Felicia LeClere, JoAnne O'Rourke, James McNally, Russell Hathaway, Kristine Witkowski, Kelly Zidar, Tannaz Sabet, Lisa Quist, Robert Melendez, and Jeremy Albright for their contributions to the LEADS database at ICPSR. The creation of the database was also supported by the following research projects at ICPSR: P01 HD045753, U24 HD048404, and P30 AG004590.

7. REFERENCES

[1] Alpert, H. 1955. The Social Sciences and the National Science Foundation. Proceedings of the American Philosophical Society, 99(5), Conference on the History, Philosophy, and the Sociology of Science: 332-333.

[2] Alpert, H. 1960. The Government's Growing Recognition of Social Science. Annals of the American Academy of Political and Social Science, 327, Perspectives on Government and Science: 55-67.

[3] Anderson, R. G., Greene, W. H., McCullough, B. D. and Vinod, H. D. 2005. The Role of Data and Program Code Archives in the Future of Economic Research. The Federal Bank of St. Louis Working Paper Series.

[4] Bachrach, C. 1984. Contraceptive Practice among American Women, 1973-1982. *Family Planning Perspectives.* 16, 253-259.

[5] Bailar, J. C., 2003. The Role of Data Access in Scientific Replication. Paper presented at Access to Research Data: Risks and Opportunities. Committee on National Statistics, National Academy of Sciences.

[6] Berns, K. I., Bond, E. C., and Manning, F. J. (eds). 1996. Resource Sharing in Biomedical Research. Committee on Resource Sharing in Biomedical Research, Division of Health Sciences Policy, Institute of Medicine. Washington, D.C.: National Academy Press.

[7] Fienberg, S. E. (1994). Sharing Statistical Data in the Biomedical and Health Sciences: Ethical, Institutional, Legal, and Professional Dimensions. *Annual Review of Public Health*. 15, 1-18.

[8] Freese, J. 2006. Replication Standards for Quantitative Social Science: Why Not Sociology? Unpublished manuscript, University of Wisconsin-Madison.

[9] Glenditsch, N. P., Metelits, C. and Strand, H. 2003. Posting Your Data: Will You be Scooped or Will You Be Famous? *International Studies Perspectives.* 4, 89-95.

[10] Jacobs, J. A., and Humphrey, C. (2004). "Preserving Research Data." *Communications of the ACM*. 47, 27–29.

[11] Kalberer, Jr., J. T., 1992. When Social Science Research Competes with Biomedical Research. *Medical Anthropology Quarterly, New Series.* 6, 391-394.

[12] King, G. 2006. Publication, Publication. *Political Science & Politics*. 39, 119-25.

[13] King, G, Herrnson, P. S., Meier, K. J., Peterson, M. J., Stone, W. J., Sniderman, P. M., et al. 1995. Verification/Replication. *PS: Political Science and Politics.* 28, 443-499.

[14] Kuhn, T. 1970. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

[15] Levan, S. G. and Stephan, P. E.. 1991. Research Productivity Over the Life Cycle: Evidence for Academic Scientists. *American Economic Review*. 81, 114-132.

[16] Long, J. St. 1978. Productivity and Academic Position in the Scientific Career. *American Sociological Review*. 43, 6, 889-908.

[17] Louis, K. S., Jones, L. M., and Campbell. E. G. 2002. Sharing in Science. *American Scientist*, 90, 4, 304-307.

[18] Matson, J. L., Gouvier, W. D., and Manikam, R. 1989. Publication Counts and Scholastic productivity: Comment on Howard, Cole and Maxwell. *American Psychologist*, 737-739.

[19] National Institutes of Health (NIH). 2003. Final Statement on Sharing Research Data. February 26, 2003. Retrieved September 6, 2006 from http://grants.nih.gov/grants/policy/data_sharing/

[20] National Science Foundation Directorate for Social, Behavioral, and Economic Sciences (NSF-SBE). (n.d.) Data Archiving Policy. Retrieved August 21, 2006 from www.nsf.gov/sbe/ses/common.

[21] Robbin, A. 2001. The Loss of Personal Privacy and Its Consequences for Social Research. *Journal of Government Information*. 28, 5, 493-527.

[22] Sobal, J. 1981. Teaching with Secondary Data. *Teaching Sociology*. 8, 2, 149-170.

[23] Stanley, B. and Stanley, M. 1988. Data Sharing: The Primary Researcher's Perspective. *Law and Human Behavior*. 12, 2, 173-180.

UPData

A Data Curation Experiment at U.Porto using DSpace

João Rocha da Silva
Faculdade de Engenharia da Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto PORTUGAL
joaorosilva@gmail.com

Cristina Ribeiro
DEI — Faculdade de Engenharia da Universidade do Porto / INESC Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto PORTUGAL
mcr@fe.up.pt

João Correia Lopes
DEI — Faculdade de Engenharia da Universidade do Porto / INESC Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto PORTUGAL
jlopes@fe.up.pt

ABSTRACT

UPData is a scientific data curation experiment currently under development at University of Porto which aims to determine the main digital preservation needs of several research groups at the university. In the course of the experiment, eight datasets have been collected from diverse scientific domains. After conducting several interviews with researchers working at U.Porto, we have concluded that from their point of view, flexible data access is the most valued capability when analysing a preservation solution and that offering such access it is the best way to involve them in the preservation workflow. We propose an extension to the DSpace repository platform to complement it with data curation capabilities. In the proposed solution, the system ingests Excel spreadsheets containing scientific data and translates them into XML documents which can then be queried via automatically generated XQuery statements. Researchers use a search webpage designed for displaying deposited data and applying various filters to it, retrieving the parts they need without having to scan each file. The collected datasets will be used as test cases for data deposit, and also to evaluate the effort required by the curation procedure.

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: On-line Information ServicesScientific Data Preservation

General Terms

Digital Preservation, Scientific Data Curation, Repositories

Keywords

Scientific Data, Preservation, Repository, DSpace Extensions, Digital Curation

1. INTRODUCTION

Nowadays, large institutions all over the world are realising the usefulness and potential of digital preservation practices when applied to scientific data. Projects such as the Data Asset Framework [4], the Edinburgh DataShare [10] or the DANS Data Archive [8] are good examples of such efforts towards better preservation of digital data assets.

It is in this context that a scientific data curation experiment named *UPData* [7] is currently being developed at the University of Porto (U.Porto). This experiment involves the university central services and a research group from the Engineering School, and has the following goals:

1. Gathering a series of heterogeneous datasets from several research domains;
2. Determining the needs of several researchers working at U.Porto and writing a use case report to document those needs;
3. Developing an extension to the *DSpace* platform [2], complementing it with scientific data management tools;
4. Depositing the gathered datasets in this extended platform and seeking feedback from previously interviewed researchers.

Building on the experiences from the Data Asset Framework, the first step of the work was to analyse the current data management reality at U.Porto. This analysis helped determine the current data preservation concerns within 13 different research groups, belonging to 7 schools within the University. The research domains are heterogenous, including Engineering, Psychology, Economics and Education Sciences.

The dataset gathering procedures included a series of interviews with the research data creators. This step was essential to ensure the correct interpretation of the supplied datasets and provided valuable insight on the potential role of a scientific data repository in ensuring the proper preservation of this kind of data. Possible improvements in backup, annotation and sharing were enumerated and prioritised. Those which were most frequently pointed out by researchers were selected as the use cases for the proposed repository extension.

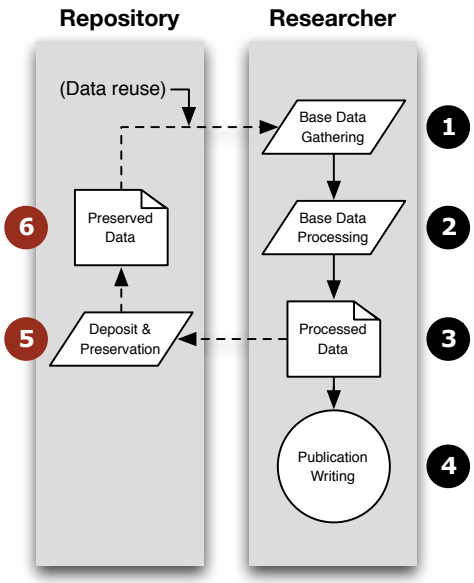


Figure 1: The research workflow and the additional data deposit steps

This solution is not intended to support the whole research workflow—it is intended to complement the publication of scientific discoveries with a data deposit and preservation step. Researchers publish their results after they complete the data gathering and analysis; however, the base data which supported the publication's findings could be better preserved, so we propose the inclusion of an additional data deposit step as shown in Figure 1. Step 1 includes the creation or gathering of base data to be processed according to the goals of the research group (Step 2). After the base data is processed (Artefact 3), conclusions are drawn and published (Step 4). We propose the inclusion of an additional data deposit step (Step 5), in which research data is gathered, annotated and translated into a preservable format (Artefact 6). Other researchers may then reuse the preserved data as base data for secondary research.

This is in line with recent policies adopted by some main scientific publications which are starting to require the inclusion of base data along with submitted articles in order to enhance the replicability of published results and provide that data to researchers in the same domain.

2. A SCIENTIFIC DATA REPOSITORY

A scientific data repository must preserve the data on a bit level and also ensure that the data is accessible and interpretable for future use. In fact, we have concluded that better accessibility is often regarded as the most interesting trait of a data repository. Researchers regard simple data backup as something they can perform on their own at a very low cost, and state that interesting data access features are important to encourage the self-deposit of data.

2.1 Use Cases

The focus on accessibility and reuse was present in many of the interviews conducted during the *UPData* experiment. Many researchers pointed out that useful repository tools should incorporate the following capabilities:

1. Easily sharing annotated datasets with their peers, reducing the need for individual follow-up contacts with researchers interested in the data;
2. Finding datasets by their dimensions, regardless of their production domain. Such dimensions are mensurable quantities or characteristics such as age, length, substances, latitude or height;
3. Exploring and querying deposited datasets through domain dimensions;
4. Retrieving query results, which involves the partial retrieval of datasets.

The data representation formats used by researchers are, in most cases, not suitable for direct data retrieval and querying. As a consequence, traditional approaches such as saving whole files and associated metadata are ill-suited for this purpose. Finding a way to reduce the granularity of data beyond the file level is a pre-requisite for building automated data manipulation capabilities.

2.2 A curation step in the data deposit workflow

There are several open-source repository solutions currently available. For this experiment, DSpace was the selected platform because U.Porto already has two operational public repositories built using this solution—the Open Repository and the Thematic Repository [11]. Contributing to DSpace can also help raise awareness on the topic of scientific data curation since the platform already has a large user base, with more than 1000 running instances [3] mainly at educational institutions. It is also open-source software, meaning that contributions can be submitted to the developer community and integrate future releases.

As part of the UPData experiment, we are designing and implementing an extension for the DSpace repository platform. This extension aims to provide users with the most requested data preservation features—easy data sharing, better searching and sub-dataset querying.

After analysing datasets gathered during the course of the experiment, we have concluded that researchers use many different formats for storing their data, which makes it difficult to develop tools to automate its processing. We have also determined that the main cause of data loss is the common lack of annotation and the use of proprietary file formats. The analysed data has, in general, quite simple models and multidimensional or hierarchical data are not very common. Most scientific data can therefore be organised into tables because the most prevalent types of files are spreadsheets, text files or other formats which can be converted into such formats by the original programs used to create the data. For these reasons, creating a better way

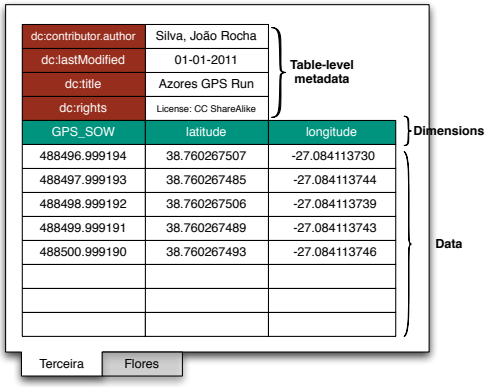


Figure 2: Example Excel spreadsheet layout

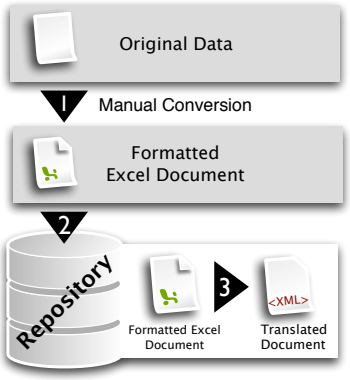


Figure 3: Data ingestion process

to manage tables in a repository platform was considered a good starting point towards better preservation of scientific data.

To make it possible for the repository to extract the relevant information from a dataset, we are designing a system which ingests specially formatted Excel spreadsheets to facilitate the interaction between the repository and the end users (either data curators or researchers). We decided to adopt this format because Excel is a common format among researchers at U.Porto and also because the implementation of a dedicated web-based deposit interface would mean heavier implementation efforts. A sample layout for a data deposit spreadsheet is shown in Figure 2. The spreadsheet is to be filled in manually by the repository curators, starting from the data which must be previously self-deposited by the original creator in its original format and layout. Since the data annotation process requires specific domain knowledge, it must be conducted in strict cooperation with the data creator. If an original data file contains several conceptual tables, each must be placed in a separate sheet of the Excel document—in this example, these are labeled "Terceira" and "Flores".

The dataset deposit and annotation workflow is depicted in Figure 3 and includes the following steps:

1. Gathering of the research data in a table-oriented format and filling in the dataset submission spreadsheet by specifying the appropriate header columns; Filling in table-level metadata, and pasting the data values—this process must be carried out by a data curator in strict cooperation with the dataset creator;

2. Submitting the spreadsheet to the repository system via the dataset ingestion page;

3. The system analyses the uploaded spreadsheet, processing each sheet for table-level metadata and column headers. Then, it matches the metadata fields with those parametrized in the repository and converts the data into an XML Document which is saved in the core DSpace database.

Since it not viable to develop automated conversion tools for all types of dataset representations, Step 1 is necessarily a manual process.

2.3 DSpace Data Explorer extension

DSpace includes a workflow engine designed to support *item* self-deposit by researchers. This workflows supports the necessary steps for the upload of dataset files and also the inclusion of relevant metadata. Such annotation can be carried out through qualified Dublin Core elements as well as other elements from additional metadata namespaces which can be parametrized in DSpace.

In DSpace terms, `Item`s are the smallest annotatable elements. These include a series of `Bitstream`s—the files contained in the `Item`. Newly submitted `Item`s must be assigned to a `Collection`. Finally, for each `Collection` there must be a group of system users, or `ePerson`s which are responsible for validating submitted `Item`s before they are published in the repository [1]—a dataset `Curator` must be a member of this group.

The deposit and indexing of datasets pose several challenges to the DSpace platform. Since dataset tables can have many different structures depending on their domain subject, a conventional relational model for such a heterogeneous reality might probably resemble the Associative Model of Data [6], with clear performance issues. XML Documents, on the other hand, have the required flexibility to represent all these different table formats and can also be queried through the XQuery language.

The high-level architecture of the DSpace extension is depicted in Figure 4, and is made up of the following modules:

1. The ingestion page can be accessed through the item viewing page in DSpace. Collection curators can upload a single formatted spreadsheet representing the data content of each of the files that make up the deposited item.

2. The XML Manager module takes care of all the operations on the XML-represented data. These include the

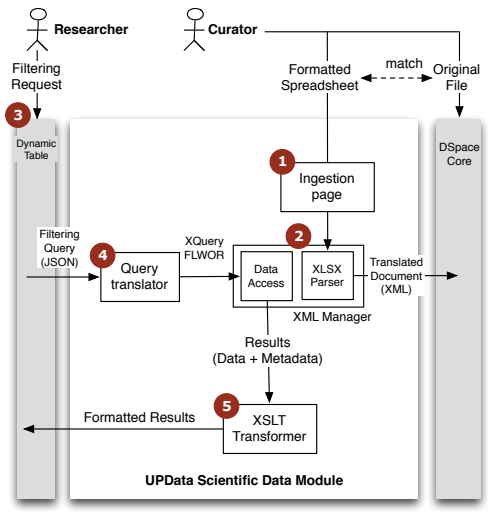


Figure 4: The UPData module architecture

translation of specially formatted spreadsheets—using the *Apache POI library* [9] for manipulating the Microsoft OOXML[1] format—and running XQuery FLWOR[2] statements over deposited data to select parts of the dataset.

3. When the user interacts with the dynamic table, a filtering statement (in the form of a JSON[3] request) is sent to the server. These filtering statements contain the column to filter by, the operator to be applied and the argument value, and can be combined using `OR`/`AND` operators to make up more complex queries. The server must then implement the required business logic to filter the data. In this case, the server side querying logic generates XQuery statements to be executed over the XML data stored in the repository.

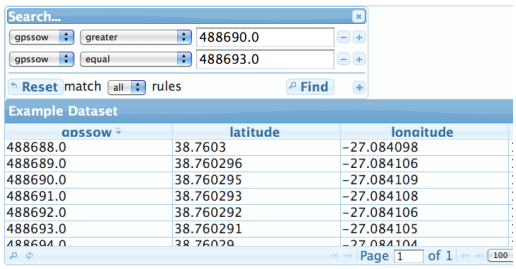


Figure 5: Web interface for a dataset table

4. The Dynamic Table component presents an XML document to the user in the form of an interactive table which is generated by the *jqGrid* library [5]. It supports basic data manipulation functionality, such as ordering data rows by specific columns in the dataset and also more complex column filtering features (numeric and string-based operators). An example table generated by the developed DSpace extension using this library is shown in Figure 5.

5. The XSLT transformer module was created to provide flexible means for presenting the data stored in the repository as well as the results of data selection. At the present time, it is used to transform the preserved data (which is stored in a rich format, complete with the relevant metadata) into a generic XML format which the *jqGrid* Javascript library can understand, so that it can create the dynamic tables shown to the repository users (see Figure 5). In the future, other transformation scenarios can be added, such as data exporting in XML-based formats or metadata-only exporting features.

[1]Office Open XML
[2]For, Let, Where, Order By, Return
[3]JavaScript Object Notation

3. ONGOING WORK

One of the planned objectives for this experiment is to obtain a reasonable estimate of the effort involved in the curation of an individual dataset. These estimates may prove to be valuable insight when considering the implementation of such practices in academic and research institutions and are to be determined in the course of this work.

4. REFERENCES

[1] P. Dietz. *DSpace 1.7.1 - System Documentation.* Duraspace, 2011.
[2] Duraspace. About DSpace. http://www.dspace.org/about.
[3] Duraspace. DSpace Registry. http://www.dspace.org/whos-using-dspace.
[4] HATII, University of Glasgow. Data Asset Framework. http://www.data-audit.eu/index.html.
[5] jQuery Grid Plugin. jQuery Grid Plugin - Grid plugin. http://www.trirand.com/blog/.
[6] T. A. Model. The Associative Model of Data White Paper technology Lazysoft. *Bookseller*, 2003.
[7] J. Rocha, C. Ribeiro, and J. C. Lopes. UPData - Scientific Data Curation at U.Porto. http://joaorosilva.no-ip.org/updata/wiki/doku.php.
[8] Royal Netherlands Academy of Arts and Sciences and Netherlands Organisation for Scientific Research. Data Archiving and Networking Services - About DANS. http://www.dans.knaw.nl/en/content/about-dans.
[9] The Apache Software Foundation. Apache POI - the Java API for Microsoft Documents. http://poi.apache.org/.
[10] University of Edimburgh. What is Edinburgh Datashare? http://datashare.is.ed.ac.uk/.
[11] University of Porto. Open Repository and Thematic Repository - repositorio.up.pt. http://repositorio.up.pt/repos.html.

Towards the Preservation of Scientific Workflows

David De Roure
Oxford e-Research Centre
University of Oxford
Oxford, UK
david.deroure@oerc.ox.ac.uk

Khalid Belhajjame
School of Computer Science
University of Manchester
Manchester, UK
khalidb@cs.man.ac.uk

Paolo Missier
School of Computing Science
Newcastle University
Newcastle upon Tyne, UK
paolo.missier@ncl.ac.uk

José Manuel Gómez-Pérez
iSOCO
Madrid, Spain
jmgomez@isoco.com

Raúl Palma
Poznań Supercomputing and Networking Center
Poznań, Poland
rpalma@man.poznan.pl

José Enrique Ruiz
Instituto de Astrofísica de Andalucía
Granada, Spain
jer@iaa.es

Kristina Hettne & Marco Roos
Leiden University Medical Center, Leiden, NL
{k.m.hettne,m.roos}@lumc.nl

Graham Klyne
Department of Zoology
University of Oxford
Oxford, UK
graham.klyne@zoo.ox.ac.uk

Carole Goble
School of Computer Science
University of Manchester
Manchester, UK
carole.goble@manchester.ac.uk

ABSTRACT

Some of the shared digital artefacts of digital research are *executable* in the sense that they describe an automated process which generates results. One example is the computational *scientific workflow* which is used to conduct automated data analysis, predictions and validations. We describe preservation challenges of scientific workflows, and suggest a framework to discuss the reproducibility of workflow results. We describe curation techniques that can be used to avoid the 'workflow decay' that occurs when steps of the workflow are vulnerable to external change. Our approach makes extensive use of provenance information and also considers aggregate structures called *Research Objects* as a means for promoting workflow preservation.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Data sharing; H.5.3 [**Group and Organization Interfaces**]: Collaborative computing

1. INTRODUCTION

Research is being conducted in an increasingly digital and online environment. Consequently we are seeing the emergence of new digital artefacts. In some respects these objects can be regarded as data; however some warrant particular attention, such as when the object includes a description of some part of the research method that is captured as a computational process. Processes encapsulate the knowledge related to the generation, (re)use and general transformation of data in experimental sciences. For example, an object might contain raw data, the description of a computational analysis process and the results of executing that process, thus offering the capability to reproduce and reuse the research process. Processes are key to the understanding and evolution of science; consequently as the scientific community needs to curate and preserve data, so we should preserve and curate associated processes [5]. The problem, as observed by Donoho et al, is that "current computational science practice does not generate routinely verifiable knowledge" [3].

In this paper we focus on computational *scientific workflows* which are increasingly becoming part of the scholarly knowledge cycle [11]. A computational scientific workflow is a precise, executable description of a scientific procedure – a multi-step process to coordinate multiple components and tasks, like a script. Each task represents the execution of a computational process, such as running a program, submitting a query to a database, submitting a job to a computational facility, or invoking a service over the Web to use a remote resource. Data output from one task is consumed by subsequent tasks according to a predefined graph topology that orchestrates the flow of data. The components (the dataset, service, facility or code) might be local and hosted along with the workflow, or remote (public repositories hosted by third parties) [9].

Workflows have become an important tool in many areas, notably in the Life Sciences where tools like Taverna [7] are popular. From a researcher's standpoint, workflows are a transparent means for encoding an *in silico* scientific method that supports reproducible science and the sharing and replicating of best-of-practice and know-how through reuse.

However, the preservation of scientific methods in the form of computational workflows faces challenges which deal precisely with their executable aspects and their vulnerability to the volatility of the resources – data and services – required for their execution. Changes made by third parties to the workflow components may lead to a *decay* of the abil-

ity of the workflow to be executed and consequently hinder the repeatability and reproducibility of their results.

This paper highlights such challenges and states the prominent role of information quality evaluation and curation in order to diagnose and react to workflow decay. Although we draw on our specific experience with workflows, the framework in this paper is designed for a more generalised notion of executable objects which we refer to as *Research Objects* [1].

We begin by discussing the difficulties underlying scientific workflow preservation (Sec. 2). We go on to highlight the role that Research Objects, as artefacts that bundle workflows together with other resources, can play in ensuring the preservation of scientific workflows (Sec. 3). We close the paper by discussing our ongoing work (Sec. 4).

2. PRESERVATION CHALLENGES

To illustrate preservation needs in scientific workflows, we use an example workflow from the field of astronomy, which is used to extract a list of companion galaxies. The workflow is illustrated in Figure 1. It starts by running two activities in parallel: the first extracts a list of companion galaxies by querying the public Virtual Observatory (VO) database, and the second activity extracts a second list of companion galaxies by invoking a web service. The results of the two activities are then cross matched to obtain an improved list of companion galaxies.

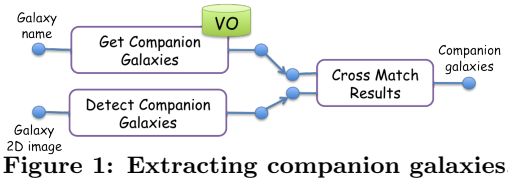


Figure 1: Extracting companion galaxies.

The content of the VO database is subject to update, and the implementation of the web service responsible for detecting companion galaxies is also subject to modifications. Thus it is possible, and likely, that the workflow produces different lists of companion galaxies when run at different times. It is important therefore to record the provenance of workflow outputs; i.e. the sources of information and processes involved in producing a particular list [12].

Should the VO database become unavailable or alter its interface so that the workflow can no longer access it, the workflow will become inoperable. This **workflow decay** is a fundamental challenge for the preservation of scientific workflows. Even though a workflow description remains unchanged, and may still have value in helping interpret results, the execution of that workflow may fail or yield different results. This is due to dependencies on resources outside the immediate context of the object which are subject to independent change. Further use cases can be found in [13] for bioinformatics and [14] for astronomy.

Gil *et al* observe that "It must be possible to re-execute workflows many years later and obtain the same results. This requirement poses challenges in terms of creating a stable layer of abstraction over a rapidly evolving infrastructure while providing the flexibility needed to address evolving requirements and applications and to support new capabilities" [4].

This abstraction approach insulates from some change, but we will still experience decay when the execution is dependent on resources and services that use independently controlled resources. For example, service providers such as the European Bioinformatics Institute (EMBL-EBI) routinely update their service offerings, and must do so in response to developments in the field of life science. Resources become obsolete or are no longer sustained. Even workflows that depend on local components are still vulnerable to changes in operating systems, data management sustainability and access to computational infrastructure. We note that workflows have many of the properties of software, such as the composition of components with external dependencies, and hence some aspects of software preservation [10] are applicable. We also observe that the above requirement is actually quite stringent: there may be other ways of usefully reproducing the experiment which do not rely on the identical workflow producing identical results.

We need a means to (i) evaluate the current status of the resources upon which the workflow depends and (ii) react to any signs of diagnosed decay in order to ensure workflow execution. In the Wf4Ever project[1] we are addressing this twofold goal through the combination of techniques for computing **information quality** and, more specifically, the integrity and authenticity of the associated resources, and **curation** techniques. Foreseeing the case where actual reproducibility cannot be achieved despite such efforts, we propose **partial reproducibility** as the means required to *play back* workflow execution based on the provenance of previous executions.

2.1 Reproducibility in scientific workflows

To provide a framework for this discussion we briefly analyse, in abstract terms, the key scenarios that arise when attempting to reproduce a workflow execution. The short formalism that follows identifies four cases for consideration here and can readily be used to discuss other cases.

Let $W_{S,D}$ denote a workflow W with dependencies on a set of services S and on a data state D. A typical example would be a bioinformatics workflow that depends on a set S of EBI services, some of which provide query capabilities into some of the EBI databases. Here D represents the content of those databases. Let $exec(W_{S,D}, d, t)$ denote the execution of W on input dataset d at time t.

As noted earlier, both service specification and implementation will evolve over time (and some services may be retired), and the state of the databases will change as well. Let S' and D' denote the new service and data dependencies at some later time t' (possibly months, or years). At this time, an investigator may be interested in using W with the following goals, and corresponding concrete options:

1. *Updated workflow on original data.* To update the old outcomes using the current, updated state of services and databases (possibly, to compare with the original outcomes): $exec(W_{S',D'}, d, t')$.
2. *Updated workflow on new data.* To test the workflow in its current state on a new dataset: $exec(W_{S',D'}, d', t')$.
3. *Original workflow on new data.* To replicate the original experiment on a new dataset d': $exec(W_{S,D}, d', t')$.
4. *Original workflow on original data.* To confirm earlier claims on the original outcomes. This translates into $exec(W_{S,D}, d, t')$, i.e., the same input d is used on W's original configuration.

[1] http://www.wf4ever-project.org

Different issues arise in each of these four cases. Cases (1) and (2) highlight workflow decay, primarily due to the evolution $S \rightarrow S'$. This is a difficult problem, which involves the evaluation of the integrity and authenticity of S and D as they evolve into S' and D' respectively and some form of ongoing curation of W in order to make it compatible with S'. We describe three approaches to curation below (Sec. 3.1), amongst which the first two have been investigated in the context of the myExperiment workflow repository [2]. For information quality, we propose provenance as an important type of evidence that can support the detection of workflow decay with respect to external resources S and D (Sec. 2.3). By providing scientists with such provenance-enabled diagnosis, we aim at feeding curation systems with accurate information of what is causing workflow decay, how and why.

Cases (3) and (4) are increasingly relevant in e-Science, as they are paradigmatic of the emerging *executable publications* [8] scenario. In an executable paper, some of the quantitative results (tables, charts) that appear in the publication are not a static part of the text, but rather they are dynamically linked to the process that produced them. In our case, the results $exec(W_{S,D}, d, t')$ are published in the paper, but they are also linked to $W_{S,D}$ as well as the input d. The intent of this emerging form of "active publication" is precisely to let readers replicate, entirely or in part, the computational portion of an experiment in order to reproduce its results. For example, Koop *et al.* [8] proposed a method that automatically captures provenance information of the experiments in order to assist authors by integrating and updating experiment results into the paper as they write it.

Supporting this scenario is not simple as it requires the entire set of original resources S and D, to be available at time t', along with the guarantee that a suitable runtime environment can be provided for the services, as well as any other software component in S. Although approaches based on Virtual Machines (VM) are common in this case [6], the high volume of state data, along with third party services that cannot be replicated locally, and the potentially high cost of execution for computationally expensive workflows, may make this approach infeasible. For example, modelling 3D data of galaxies [14] involves the manipulation of large data cubes, the size of which may reach tens of terabytes. *Partial reproducibility* alleviates the problem in practical cases.

2.2 Partial Reproducibility

Consider the astronomy workflow presented in Figure 1. For (3) and (4), insisting on executing W in its original environment is not always feasible and may not be needed. An executable paper may for example provide limited workflow execution capability to readers, permitting only execution of lightweight tasks, such as analysis and charting of tabular data, as opposed to compute-intensive simulations. This corresponds to splitting the workflow into two portions (top/bottom), where only the latter is made available for readers to experiment with, while they will still have to rely upon the usual peer-review guarantees regarding the correctness of the top portion of the workflow.

Executing W is unnecessary provided that a complete and reliable provenance trace has been recorded at time t. By combining provenance traces with partitioned executable workflow fragments, provenance can be used to "play back" the original execution and can be queried to inspect all data dependencies that resulted from that execution: *(i)* the provenance is recorded from the execution of fragments that are heavily dependent on S and D, which are then omitted, and *(ii)* a VM approach is used for the remaining segments, which are executable. Partitioning requires that the executable segments be found downstream (in terms of the directed graph that represents the workflow structure) from the omitted fragments. This places a requirement on workflow design. Minimising the associated *cost* to reproducibility of a workflow, which involves S, D, and the actual cost of execution (which may well be a monetary cost, for example in the case of cloud-based computations) presents the challenge of finding the optimal partitioning. These are just two of the challenges arising from taking this pragmatic approach.

2.3 Information quality evaluation

In order to detect workflow decay with respect to the evolution of the tuple (S,D) of services and data needed for workflow execution, we focus on two main aspects relevant for information quality: integrity and authenticity. *Integrity* refers to the quality or condition of being whole, complete and unaltered while *authenticity* addresses the lineage of data.

One of the main sources required to evaluate information quality is provenance information (in our case about S and D) which offers the means to verify the evolution of data and services, to analyse the processes that led to their current status, and to decide whether they are still consistent with a given workflow. We build on provenance to compute the integrity and authenticity of workflows with respect to (S,D), thus providing scientists with accurate information about what is causing the workflow decay due to changes in such resources, how and why.

We can use and extend existing provenance vocabularies like the Open Provenance Model[2] to record and reason about provenance metadata relevant to the diagnosis of workflow decay. Additional challenges include providing scientists with the means to interpret easily the results of such analysis and to assist them in the early diagnosis of workflow decay and the selection of the most appropriate curation techniques.

3. PRESERVING WORKFLOWS USING RESEARCH OBJECTS

3.1 Preservation in Practice

The myExperiment[3] social website for finding, storing and sharing workflows has been in operation since 2007 and holds the largest public collection of scientific workflows [2]. As such it provides a useful case study in workflow decay and preservation, supporting two main mechanisms.

First, the continual downloading and uploading of workflows provides a *community curation* mechanism for workflows that are reused, and these in turn can act as examples to inform community members when updating other workflows. Expert curators, e.g. scientists, are involved in annotating workflows, by tagging and providing exemplars.

The second mechanism is *assistive curation* using semi-automated processes to perform 'housekeeping' on the corpus of workflows. For example, when a service provider an-

[2]http://openprovenance.org
[3]http://www.myexperiment.org

nounces that a service is deprecated and will be removed or replaced on a certain date, the workflows affected by this can be tagged accordingly and replacement advice propagated to the appropriate users. Potentially this could progress to *autonomic curation* where workflows could be executed and repaired automatically, for example when services change.

The assistive approach keeps the 'human in the loop' and the Wf4Ever project is pursuing this by focusing on *recommendations* for curation and repair; for example a replacement for a service can be confirmed using provenance logs.

3.2 Research Objects

Workflow specifications are insufficient for guaranteeing the preservation of scientific workflows. The reproducibility strategies listed in Sec. 2.1 show that, in addition to workflow specification, we need information about the components that implement workflow steps, the data used and produced as a result of workflow enactment.

In practice myExperiment users sometimes choose to aggregate workflows with associated data (in 'packs') and this provides a powerful means to track S and D. Building on packs, to cater for workflow preservation we use the notion of a *Research Object*, which can be viewed as an aggregation of resources that bundles workflow specification and additional auxiliary resources. These may include input and output data which enables workflows to be validated.

The elements that compose a Research Object may differ from one to another, and this difference may have consequences on the *level* of reproducibility that can be guaranteed. At one end of the spectrum, the Research Object is represented by a paper. As we progress to the other end the Research Object is enriched to include elements such as the workflow implementing the computation, annotations describing the experiment implemented and the hypothesis investigated, and provenance traces of past executions of the workflow. Assessing the reproducibility of computations described using electronic papers can be tedious: a paper may just sketch the method implemented by the computation in question, without delving into details that are necessary to check that the results obtained, or claimed, in the paper can be reproduced. Verifying the reproducibility of Research Objects at the other end of the spectrum is less difficult. The provenance trace provides data examples to re-enact the workflow and a means to verify that the results of workflow executions are comparable with prior results.

To ensure the preservation of a workflow and the reproducibility of its results, the Research Object needs to be managed and curated throughout the lifecycle of the associated workflow. The provenance of the Research Object elements (i.e., workflow, data sets and web services) is key to understanding, comparing and debugging scientific workflows and to verifying the validity of a claim made within the context of a Research Object by revealing the data inputs used to yield a given workflow result. We need to support the logging, browsing and querying of the provenance linking components of Research Objects and the traces of workflow executions.

4. CONCLUSIONS

As research practice evolves we anticipate a growing quantity and diversity of executable objects, in particular computational scientific workflows. We outlined the challenges underlying the preservation of scientific workflows and sketched preliminary solutions that can be adopted for that purpose. We used the concept of Research Object as an abstraction for the management of executable objects throughout their life-cycle. We anticipate that this work will give rise to recommendations and best practices for authors and curators of scientific workflows to meet preservation requirements.

We are investigating the reproducibility and curation strategies reported in this paper and developing a software architecture and reference implementation for workflow preservation. The development of the reference implementation will rest on existing developments in scientific workflow repositories, digital libraries and preservation systems. In particular, we will build on well-established digital libraries, such as dLibra[4], to extend the myExperiment workflow repository with further preservation capabilities.

5. ACKNOWLEDGMENTS

Wf4Ever is funded by the Seventh Framework Programme of the European Commission (Digital Libraries and Digital Preservation area ICT-2009.4.1 project reference 270192). myExperiment is funded by UK JISC. The dLibra Digital Library Framework has been produced by the Poznań Supercomputing and Networking Center since 1999. We are grateful to all our collaborators in these projects.

6. REFERENCES

[1] S. Bechhofer, I. Buchan, D. D. Roure, et al. Why linked data is not enough for scientists. *Future Generation Computer Systems*, doi 10.1016/j.future.2011.08.004, 2011.

[2] D. De Roure, C. Goble, and R. Stevens. The design and realisation of the myExperiment virtual research environment for social sharing of workflows. *Future Generation Computer Systems*, 25(5):561–567, 2009.

[3] D. L. Donoho, A. Maleki, I. Rahman, et al. Reproducible research in computational harmonic analysis. *Computing in Science and Engg.*, 11:8–18, January 2009.

[4] Y. Gil, E. Deelman, et al. Examining the challenges of scientific workflows. *IEEE Computer*, 40:24–32, Dec. 2007.

[5] C. Goble and D. De Roure. Curating scientific web services and workflows. *Educause Review*, 43(5), 2008.

[6] P. J. Guo and D. Engler. CDE: Using System Call Interposition to Automatically Create Portable Software Packages. In *Proc. USENIX Annual Tech. Conf.*, 2011.

[7] D. Hull, K. Wolstencroft, R. Stevens, et al. Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34(suppl 2):W729–W732, 1 July 2006.

[8] D. Koop et al. A provenance-based infrastructure to support the life cycle of executable papers. *Procedia Computer Science*, 4:648 – 657, 2011. Proceedings of the International Conference on Computational Science.

[9] B. Ludäscher et al. Scientific process automation and workflow management. In *Scientific Data Management*, Computational Science Series. Chapman & Hall, 2009.

[10] B. Matthews et al. A framework for software preservation. *International Journal of Digital Curation*, 5(1), 2010.

[11] J. P. Mesirov. Accessible reproducible research. *Science*, 327(5964):415–416, 2010.

[12] P. Missier. *Modelling and computing the quality of information in e-science*. PhD thesis, University of Manchester, 2008.

[13] M. Roos. Genomics Workflow Preservation Requirements. Technical report, Deliverable D6.1, Wf4Ever project, 2011.

[14] L. Verdes-Montenegro. Astronomy Workflow Preservation Requirements. Technical report, Deliverable 5.1, Wf4Ever project, 2011.

[4] http://dlibra.psnc.pl

People Mashing: Agile Digital Preservation and the AQuA Project

Paul Wheatley
British Library
Boston Spa
Wetherby
+44(0)1937546254
paul.wheatley@bl.uk

Bo Middleton
Brotherton Library
University of Leeds
Leeds
+44(0)1133436386
m.m.middleton@leeds.ac.uk

Jodie Double
Brotherton Library
University of Leeds
Leeds
+44(0)1133437783
j.l.double@leeds.ac.uk

Andrew N. Jackson
British Library
Boston Spa
Wetherby
+44(0)1937546254
andrew.jackson@bl.uk

Rebecca McGuinness
Open Planets Foundation
Boston Spa
Wetherby
+44(0)1937546254
rebecca@
openplanetsfoundation.org

ABSTRACT

Manual quality assurance (QA) of digitised content is typically fallible and can result in collections that are marred by a variety of quality and access issues. Poor storage conditions, technology obsolescence and other unforeseen problems can also leave digital objects in an unusable state. Detecting, identifying and ultimately fixing these issues typically requires costly and time consuming manual processes. An inadequate understanding of potential tools and their application creates a barrier to the automation and embedding of preservation processes for many collection owners. The JISC funded [1] Automating Quality Assurance Project (AQuA) [2] applied a variety of existing tools in order to automatically detect quality and preservation issues in digital collections and work to bridge the divide between technical and collection management expertise. Two AQuA Mashup events brought together digital preservation practitioners, collection curators and technical experts to present problematic digital collections, articulate requirements for their assessment, and then apply tools to automate the detection and identification of the content issues. By breaking down the barriers between technical and non-technical practitioners, the events enabled grass-roots digital preservation collaboration between the two communities. This paper describes the AQuA Project's novel approach to agile preservation problem solving and discusses the incidental benefits and community building that this strategy facilitated.

1. THE CHALLENGE

Creating a digital object via digitisation is prone to mistakes and the introduction of quality issues. In recent years, increasingly ambitious digitisation programmes (such as the recent JISC eContent Programme [3]) have turned digital content creation into a mass production activity. Known quality issues include missing pages, duplicate pages, incorrect de-skew, out of focus images, incorrect or incomplete metadata, the infamous "thumb in picture" and a variety of other processing or corruption problems have been introduced with mass digitisation.. (see Figure 1). Collection curators and technical staff are now faced with detecting mistakes and quality issues on a large and ever expanding scale.

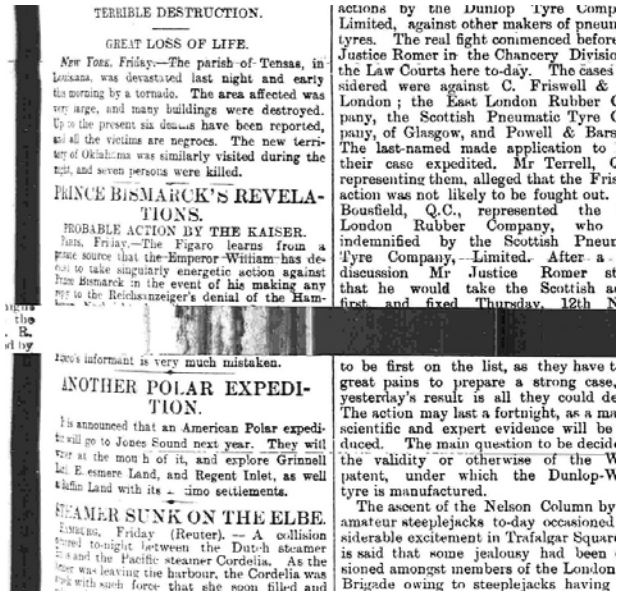
Undetected digitisation quality issues can become digital preservation issues later in the lifecycle and these are often problems that are hard to rectify once the source material has been re-shelved and the digitisation activity has been closed. With only manual content checking to mitigate these issues, there is a serious risk of erroneous or poor quality content making it through to the end user's screen. Timely and automated identification of problematic scans would enable re-digitisation at comparatively low cost as opposed to costly retrospective rescanning years later.

Preserving an existing digital object (whether digitised or born digital) typically requires a number of processing steps, before it can be safely placed into a digital repository. Each of these individual operations has the potential to malfunction, sometimes with disastrous results for the resulting preservation effort. Whenever digital content is acquired, created, moved, un-packaged, processed, migrated, curated, repackaged or otherwise changed, problems can occur and collection damage can result. Culprits include software bugs, network dropouts, full disks and human error.

Detecting these issues requires thorough content checking at key lifecycle stages. File hashing and file manifests can support efficient digital object integrity checking, but many operations in a preservation workflow will legitimately alter the digital objects, resulting in a necessary recalculation of file hashes. Manual checking of content is a typical method of catching systematic errors, but suffers from a number of drawbacks. Human effort can be costly and this makes it difficult to scale this approach up to support the QA of large collections. A visual check can sometimes be subjective and QA problems can be quite subtle and hidden. Sampling approaches can also be used, but this leaves

iPRES2011, Nov. 1–4, 2011, Singapore.

blind-spots where issues can remain undetected. A more thorough and automated QA check may prove to be unaffordable unless built into core business practice for the collection managers and their institutions.



TERRIBLE DESTRUCTION.

GREAT LOSS OF LIFE.

New York, Friday.—The parish of Tensas, in Louisana, was devastated last night and early this morning by a tornado. The area affected was very large, and many buildings were destroyed. Up to the present six deaths have been reported, and all the victims are negroes. The new territory of Oklahoma was similarly visited during the night, and seven persons were killed.

PRINCE BISMARCK'S REVELATIONS.

PROBABLE ACTION BY THE KAISER.

Paris, Friday.—The Figaro learns from a private source that the Emperor William has decided to take singularly energetic action against Prince Bismarck in the event of his making any reply to the Reichsanzeiger's denial of the Ham-

actions by the Dunlop Tyre Comp Limited, against other makers of pneu tyres. The real fight commenced before Justice Romer in the Chancery Divisio the Law Courts here to-day. The cases sidered were against C. Friswell & London ; the East London Rubber pany, the Scottish Pneumatic Tyre pany, of Glasgow, and Powell & Bars The last-named made application to their case expedited. Mr Terrell, representing them, alleged that the Fris action was not likely to be fought out. Bousfield, Q.C., represented the London Rubber Company, who indemnified by the Scottish Pneu Tyre Company, Limited. After a discussion Mr Justice Romer st that he would take the Scottish a first and fixed Thursday, 12th

burg's informant is very much mistaken.

ANOTHER POLAR EXPEDITION.

It is announced that an American Polar expedition will go to Jones Sound next year. They will winter at the mouth of it, and explore Grinnell Land, E. esmere Land, and Regent Inlet, as well as Baffin Land with its Esquimo settlements.

STEAMER SUNK ON THE ELBE.

Hamburg, Friday (Reuter). — A collision occurred to-night between the Dutch steamer Holland and the Pacific steamer Cordelia. As the latter was leaving the harbour, the Cordelia was struck with such force that she soon filled and

to be first on the list, as they have t great pains to prepare a strong case, yesterday's result is all they could de The action may last a fortnight, as a ma scientific and expert evidence will be duced. The main question to be decid the validity or otherwise of the W patent, under which the Dunlop-W tyre is manufactured.

The ascent of the Nelson Column by amateur steeplejacks to-day occasioned siderable excitement in Trafalgar Squar is said that some jealousy had been sioned amongst members of the London Brigade owing to steeplejacks having

Figure 1: A portion of a digitised newspaper image exhibiting damage arising during post processing

If manipulating content increases the chance of damaging it in an unforeseen way, leaving it untouched over time raises the potential for obsolescence issues to be encountered. The critical questions facing digital preservationists include: will this content render correctly on the user's computer? Is it likely to render correctly in 5, 10 or 20 years time? If not, why not and what can be done about it? A variety of more technical proxies are typically raised in an effort to begin to answer these challenging questions. What is the file format? Does this file validate to its file format specification? Are there any external dependencies? There is therefore a need to assess or characterize digital content in order to gain a better understanding of its properties, analyze potential risks and inform subsequent preservation planning and remedial preservation treatments.

Quantifying the incidence and impact of these problems is difficult, particularly with regard to quality rather than preservation issues. The authors had encountered quality or processing problems at their respective institutions. However, organizations are usually not pro-active about broadcasting what might unfairly be seen as bad news stories. Anecdotal evidence suggested these issues were not uncommon elsewhere and documented QA work such as that by Riley and Whitsel, 2005 [4] also implies the existence of a challenge to be met. But prior to the AQuA Project, the real significance of these issues for memory and higher education institutions remained somewhat unclear as this was a collection management issue that was rarely discussed openly and more importantly discussed between technical and non-technical staff.

2. POTENTIAL SOLUTIONS

The authors felt that potential existed to apply existing software tools to many of the problems outlined above in addition to engaging collection curators in preservation planning who are normally excluded from hackathons. Several pre-requisites in terms of knowledge, access to data and expertise would need to be met for significant progress be made during the events:

1. A good understanding of the specific QA and preservation challenges faced by institutions.
2. Access to samples of problematic digital collections where these challenges were present, to support solution testing
3. Knowledge of likely toolsets that might provide useful solutions
4. Effort to progress solutions

The authors identified a potential funding stream from the Joint Information Systems Committee (JISC) that matched well with the problem space. The University of Leeds led a successful bid, partnering with the University of York, the British Library and the Open Planets Foundation. Funding conditions restricted the project length to 6 months and a modest budget. These constraints would make it difficult to gather QA and preservation problems and associated content, discover likely toolsets, apply them to the problems and evaluate the results all within the limited project length. Recruitment of project staff would be challenging due to a very short project lead time, and finding sufficient staffing expertise to meet the pre-requisites listed above nigh on impossible. The collaborators (represented by the authors of this paper) therefore pursued a more agile approach which would engage with practitioners and experts from other institutions in 2 mashup events that would each be 3 days in length. This would have the added benefits of gaining buy in to project outputs by getting potential users involved in creating the solutions, while facilitating knowledge sharing and collaboration.

3. THE AQUA MASHUP APPROACH

The AQuA approach has its origins in the Hackathon [5], where software developers meet up to solve technical challenges over a short period of time. Hackathon events have become increasingly popular in recent years as a way of removing the overhead of traditional project based development and enabling rapid prototyping and development through a combination of collaboration and friendly competition. The digital library community has begun to embrace the Hackathon concept, with projects such as DEVCSI [6], working actively to develop a technical community via supporting activities such as Hackathons and programming challenges.

The advent of the open data and linked data approaches has encouraged the creation of a similar event model to the hackathon but with a focus on exploiting open interfaces, mashing up data from several sources and providing new and often innovative services. Data Mashup [7] events, like Hackathons, typically provide supportive environments for participants to collaborate in small teams and compete to win challenges.

The Unconference [8] approach, demonstrated in the repository community by the CURATEcamp [9] events, seeks to break away from the pre-planned and often rigid structure of typical face to face meetings and support a more agile and bottom up approach.

The AQuA Project Mashups drew on elements of these existing approaches, while adding some new concepts in order to meet the challenges described above. Rather than being purely technically focused AQuA invited software developers as well as digital preservation practitioners and curation staff and gave them specific roles to play during the events. Instead of setting challenges for the attendees, we asked them to bring along issues they needed solutions to be developed for and spent time capturing and recording these in order to support future work. Although not quite a Hackathon, Mashup or Unconference, the authors settled on describing the events as Mashups.

4. THE AQUA EVENTS

The AQuA Project organized two Mashup events. The first was held at Weetwood Hall in Leeds for 18 attendees in April 2011. The second event was held at the British Library in London for 30 attendees.

4.1 Mashup Event Planning

A substantial amount of pre-event planning focused ensuring the attendees understood the expectations from the team and that the event ran smoothly.. A strict "no observers" rule required that every attendee had to either bring collection content with them and champion it at the event, or have the skills to play a developer role.

4.2 Mashup Event Format

The first day of each AQuA Mashup focused on setting the scene and capturing the digital preservation challenges that would be tackled. After a brief introduction to outline the structure of the event the focus was quickly placed on the participants, who gave lightning talks to the group. Attendees playing the role of Collection Owners were asked to bring along samples of problematic digital collections and talk about the issues they had. Technical attendees were asked to talk about their skills, experience and interests. Over lunch the facilitators matched up the attendees into teams, ensuring that each Collection Owner was supported by a Developer. Working in small groups, and in some cases individual teams, details of the collections samples brought to the event were discussed. QA and preservation issues were identified and recorded, and potential avenues to explore in solving the challenges were noted. From this brainstorm, teams were able to select a challenge they were interested in tackling and begin work on it. The Developers began to seek out useful software tools to apply in order to tackle the identified issue, while the Collection Owners recorded the results of the brainstorming and progress made with solutions.

The second day had much less structure, allowing the Developers plenty of opportunity to progress their technical work, while liaising closely with the Collection Owners on their teams. Collection Owners had the opportunity to work further on capturing their preservation issues and broadening the perspective to explore contextual challenges. Institutional constraints would inevitably impact on the technical solutions being developed and how they could ultimately be embedded into existing workflows.

The third day initially provided some time to wrap up development work, focus on capturing, and where possible visualizing, the results. A small group brainstorm was facilitated to consider the next steps once the event had concluded. Lightning talks to report back results to the group were followed by opportunities to evaluate the solutions and discuss the AQuA approach and events. Prizes for the best work by a Developer and the best work by a Collection Owner were voted on by the attendees themselves.

A strong focus was placed on capturing all event outputs on either the project wiki or Git code repository as they were developed or understood. A key concern of the authors in focusing project development effort into short lived Mashup events was that useful work might easily be lost if not captured straight away. Post event wiki gardening was planned to ensure a clear and meaningful record of results was captured and publicly available [2].

5. PROJECT RESULTS

5.1 Collections, Issues and Solutions

The AQuA Project wiki [2] contains descriptions of the outputs of the project events. Each of the digital content samples brought along to an AQuA event is listed and described under the Collections section. This described the basic details of the collection and provided a high level description of its characteristics. Preservation or QA challenges were termed "Issues" and listed under a related wiki page. These issues were related to specific collections using hyperlinks. All Issues were recorded in a standard proforma, capturing a detailed description of the preservation or QA challenge as well as possible approaches for tackling it. Where AQuA was able to explore a solution to the issue, a further "Solution" wiki page was produced. This described the approach taken and provided a link to the Solution itself and contained review notes on how well the Solution had solved the related Issue. The resulting network of Collections, Issues and Solutions provides a permanent record of the AQuA results.

5.2 People Mashing

A key aim of the project was not only to develop some solutions to the QA and preservation challenges identified, but also to facilitate collaboration, knowledge sharing and hopefully lasting relationships between the attendees.

Mahey and Walk [10] identify a need to break developers out of constrained development and problem solving cycles and exploit their wider capability while also developing them as individuals. They go on to describe how face to face events, amongst other possibilities, can facilitate collaboration, knowledge sharing and develop a support community. AQuA took this further by breaking down the barriers between technical and none-technical staff, creating an environment where participants were happy to ask questions without fear of judgement, and encouraging agile problem solving. AQuA dubbed this approach "People Mashing".

Non-technical staff developed skills to articulate issues and technical staff were able to develop preservation tools that would have an impact beyond the event. Participants commented in both a discussion at the end of the second AQuA event and in anonymous feedback that they were keen to encourage and maintain the community that the events had begun to establish.

6. REVIEW AND LESSONS LEARNT

6.1 Feedback, Review and Refinement

Survey Monkey was used to gather feedback from attendees at both events and time was made at the end of the London Mashup to discuss as a group how the event went and what the organizers and attendees should do next. Several planning and review meetings were held between events where the schedule was

revised and each session updated to take advantage of the experience of running the first event and the feedback received. Scaling up aspects of the first event to work with double the number of participants for the London Mashup was a key challenge.

6.2 What worked well

The popularity of the events and the presence of an array of both preservation and quality issues in participants' collections vindicated the project focus. Indications that these issues were actually a significant issue for many institutions were confirmed.

The events yielded a significant number of functional preservation solutions, some prototypes that required further work and a number of promising problem/solution explorations that can all be found on the wiki. Several participants were keen to stress that they would be taking home workable solutions that they could put into practice straight away. Peer review by the collection owners of the solutions developed for them was largely positive, although many noted that more development and support would be needed and illustrates a long-term challenge from the events to continue testing and refinement of tools in production environments.

Capturing a record of each Mashup using Collection/ Issue/ Solution proformas worked well in providing structure and clear aims for the events while ensuring that the valuable work performed was not lost at the end of the Mashups. The resulting documentation should be useful in supporting adoption and re-use of AQuA results by the Open Planets Foundation and other interested parties.

Many attendees gave very positive feedback about the collaborative and inclusive nature of the events. Several comments focused on the benefits of the agile approach to working. One attendee commented "Putting 30 people into a room, some with problems and some who can write solutions is extremely eye opening. I've learnt that free from restrictions on infrastructure and process … prototyping can solve a varied number of non-trivial problems quickly."

A number of the solutions developed took a genuinely innovative approach, such as the RDF visualization of characterization results [11] produced at the London Mashup. Encouraging participants to work on new problems, often outside their comfort zone, and discuss their approaches with others helped to facilitate this.

6.3 What worked less well

Collection Owners weren't challenged enough on the second day when the focus was on progressing the technical solutions. More sessions focusing on preservation planning and next steps would have made better use of their time and given them a tangible piece of work to take back to their institutions.

Following the first Mashup, it was clear that development time at the event needed to be maximized and as a result lightning talks for reporting back were minimized. This was probably a mistake as it would have increased interaction between the teams and sharing of ideas between developers.

Formal checkpoints between Developers and Collection Owners may have helped to reduce the length of development cycles, although many teams worked closely enough for this not to have been a significant issue.

Conference venues were used to host both events which precluded late night coding sessions. Several of the Developers were disappointed not to be able to keep working into the evening on the second day. Focusing the first evening on a meal and social event to encourage networking and the second as all night hack time would have been a good compromise.

Three days is also a long time for participants to abandon their day job and join a Mashup or Hackathon event. A number of interested parties would like to have joined one of the AQuA events but were unable to convince their manager to release them for the duration. On the other hand, fitting a structured event into less than three days would have been challenging. Project funding to cover accommodation and catering helped participants to justify time on AQuA as there were few additional costs to them.

Good Wi-Fi is essential at an event of this kind. Signal strength problems were encountered at the London event and a backup plan had to be put into action at short notice. Having a reserve ready to go is recommended.

7. NEXT STEPS

At the time of writing the AQuA Project Team is planning a follow up event that will focus on evaluating adoption of project results. It will consider what barriers there are to further development or re-use of the tools with the aim of targeting effort from the Open Planets Foundation, JISC and others on appropriate support activities.

Given the success of the AQuA events in beginning to build a community of digital preservation practitioners, maintaining the momentum with further face to face events would be desirable. All but one of the attendees who completed the feedback survey for the London event stated that they would like to attend more mashup events of the same AQuA format. Since the completion of the AQuA Project itself, the OPF and the Digital Preservation Coalition have announced a new event that has adopted the AQuA mashup format and approach [12].

8. REFERENCES

[1] *Grant 15/10: JISC infrastructure for education and research programme*, http://www.jisc.ac.uk/fundingopportunities/funding_calls/2010/10/grant1510.aspx

[2] *Automating Quality Assurance Project*, http://wiki.opf-labs.org/display/AQuA

[3] *JISC eContent Capital Programme*, http://www.jisc.ac.uk/fundingopportunities/funding_calls/2011/06/econtentcapital.aspx

[4] Riley, J and Whitsel, K, 2005. Practical quality control procedures for digital imaging projects, OCLC Systems & Services Volume: 21 Issue: 1. http://www.dlib.indiana.edu/~jenlrile/publications/imageqc/qc.pdf

[5] *Hackathon*. Wikipedia. http://en.wikipedia.org/wiki/Hackathon

[6] *Developer Community Supporting Innovation Project*, http://devcsi.ukoln.ac.uk/blog/about/

[7] *Mashup (digital).* Wikipedia. http://en.wikipedia.org/wiki/Mashup_%28digital%29

[8] *Unconference*, Wikipedia. http://en.wikipedia.org/wiki/Unconference

[9] *Curate Camp*, http://curatecamp.org/about

[10] Mahey, M and Walk, P. 2010. *Why UK Further and Higher Education Needs Local Software Developers.* Ariadne, Issue 65. http://www.ariadne.ac.uk/issue65/mahey-walk/

[11] Cliff, P and Fay, E. *tiff2RDF - visualising image collection consistency*. http://wiki.opf-labs.org/display/AQuA/tiff2RDF+-+visualising+image+collection+consistency

[12] *OPF and DPC Hackathon: Practical Tools for Digital Preservation*, http://www.openplanetsfoundation.org/community/opf-events/hackathon-practical-tools-digital-preservation

From the World Wide Web to Digital Library Stacks: Preserving the French Web Archives

Clément Oury
Bibliothèque nationale de France
Legal Deposit Department
clement.oury@bnf.fr

Sébastien Peyrard
Bibliothèque nationale de France
Bibliographic and Digital Information Department
sebastien.peyrard@bnf.fr

ABSTRACT

The National Library of France is mandated by French law to collect and preserve the French Internet. It is now a 10-year old project with collections ranging from 1996 to the present. To ensure their long-term preservation, the choice has been made to ingest these web archives into the institution's existing digital preservation repository, SPAR (Scalable Preservation and Archiving Repository). There were numerous implementation challenges, on the modeling as well as the technical sides, which the library met with solutions drawn from international collaboration and widely adopted standards, whenever possible.

– Web archive-specific formats (W/ARC files) lacked validation and characterization tools, which led to the development of a Jhove2 module for the ARC format.

– The heterogeneity of BnF's web archives in terms of formats, production workflows and tools, was managed by aligning all of them on a single model, the current production workflow using NetarchiveSuite.

– The specificities of web archives were matched to the PREMIS data model and dictionary and SPAR's global METS profile.

– Finally, the need to express technical information about ARC files in a concise, manageable fashion led us to define a format-specific metadata scheme for container files, containerMD, which will be released to the preservation community (on BnF's website).

All this development work means new services for digital curators in general and preservation experts in particular. They will be able to know their collection better, to check its comprehensiveness, and, with that deeper understanding, to investigate new preservation strategies. Allowing differentiated service level agreements for specific sets of documents, with richer metadata extraction, better quality insurance and differentiated preservation strategies, will be the logical next step of the web archives long-term preservation project.

Keywords

Web archives; Metadata; Characterization tools; ARC file format.

1. THE ISSUE: INGESTING THE LEGAL DEPOSIT OF THE FRENCH INTERNET IN BnF'S DIGITAL REPOSITORY

1.1 The legal deposit mandate

As of August 1st, 2006, a copyright law gives the National Library of France the mandate to collect a new kind a resource: the whole set of data that is publicly available on the French Internet. This mandate has been given to the library thanks to an extension of legal deposit, which obliges every publisher to send copies of his output to the library. The Internet having obviously become the favorite place to create and distribute knowledge and information, it was necessary to give French heritage institutions the legal means to ensure its preservation[1].

Although the law was voted in 2006, the project of collecting French websites at BnF dates back to the early years of the last decade. In 2002 was launched the collection of all websites related to the presidential and parliamentary elections; this operation was renewed two years later, for European and then regional elections. These crawls were performed with a small-scale harvesting robot, called HTTrack[2].

The library was still lacking the technical means (hardware and software), skills and experience necessary to realize large-scale crawls of the French web. This is the reason why BnF agreed on a partnership with Internet Archive (IA), a not-for-profit foundation involved in world-wide web archiving since 1996. Thanks to this agreement, five annual broad crawls (from 2004 to 2008) of the .fr domain were performed by IA and enriched BnF's scollections [4]. They were performed by Heritrix[3], a harvesting robot developed by Internet Archive and several Scandinavian libraries in the framework of the International Internet Preservation Consortium (IIPC)[4].

Along with the results of the annual .fr crawls, Internet Archive delivered to BnF large extracts of its own collections, from 1996 to 2005. These so-called historical collections were not directly crawled by Internet Archive, but given to this institution by Alexa Internet [3]. In terms of value, these early collections may be compared to the first printed books.

At the same time, BnF was building an infrastructure (technical as well as organizational) to perform in-house crawls. The library decided to use Heritrix, and started by conducting focused crawls

[1] About the legal aspects of Web archiving in France, see [2].

[2] HTTrack Website Copier website: http://www.httrack.com.

[3] Heritrix home page: http://crawler.archive.org.

[4] IIPC website: http://www.netpreserve.org.

on a continuously growing number of websites and resources (from 130 million URLs in 2007 to 350 million in 2010).

Finally, in 2010, BnF launched its first in-house .fr domain crawl. To achieve this goal, NetarchiveSuite, developed by the Royal Library of Copenhagen and the University Library of Aarhus, was added on top of Heritrix[5]. This tool helps curators manage the harvesting workflow for very broad or very frequent crawls. Successfully tested on the .fr domain crawl in 2010, NetarchiveSuite is now used for all focused and domain crawls.

Figure 1 : BnF web archives collections as of July 2010

Collection	Historical collections	First election crawls	IA crawls	In-house crawls without NetarchiveSuite	In-house crawls with Netarchive-Suite
Date	1996-2005	2002 and 2004	2004-2008	2006-2010	2010-…
Size	70 Tb	0.5 Tb	45 Tb	22 Tb	23 Tb
Operator	Alexa Internet	BnF	IA	BnF	BnF
Robot	-	HTTrack	Heritrix	Heritrix	NetarchiveSuite and Heritrix

In short, the harvesting process has been now fully internalized. Access to these web archives has been opened in BnF reading rooms. Ensuring their long-term preservation was a further step in order to achieve a complete library lifecycle, but two main issues arise in tackling this challenge:

- The size and variety of these collections make them invaluable and harder to preserve at the same time.
- BnF's existing digital repository, SPAR, was a natural choice for preserving our web archives, but some adjustments were necessary on both sides.

1.2 SPAR

Ingesting BnF's web archives in SPAR [1] is indeed an opportunity and a constraint at the same time.

The opportunity is great: the core preservation functions of the system have already been defined, developed and are up and running, which lowers implementation risks. Using the same preservation system for all the digital collections at BnF also has the benefit of being cost-efficient.

Apart from project and cost management issues, this is also clearly an opportunity from a librarian point of view. From its early stages, SPAR has been designed to manage heterogeneous digital documents with different preservation policies to be applied. It would be something of a waste not to use these features.

Finally, using a single repository solution for all kinds of digital documents in a single system seems more manageable over the long term: the distinction made between web archives and, say, born-digital acquisitions, can shift over time. Being able to manage them in a single system can make things easier later.

However, integrating the web archives with SPAR also has its constraints: there is a pre-existing data model [1], which could be adapted and enhanced, but not replaced by a new one; in addition, BnF's web archives are poorly described as there is currently no cataloguing of these collections, whereas the first ingested collections, of digitized books and audiovisual documents, are far better-known and described.

2. IMPLEMENTATION: FROM LOCAL ISSUES TO INTERNATIONAL COOPERATION

2.1 Knowing our collections: the Jhove2 modules

Before ingesting BnF's web archives, BnF digital curators should be able to know the technical characteristics of their collections and, thanks to this, what they can do with them in terms of preservation. The huge amount and variety of the harvested files, impossible to encompass directly, led us to concentrate for the moment on the container file levels, in particular the ARC file format[6], used for all the collections. It was vital to have tools that were able to validate and extract information about these files, *and* that allowed, at least, content files to be identified – and thus, multi-level file-format analysis features.

In order to achieve this goal, we decided to use the framework proposed by Jhove2[7]. However, this tool lacked an ARC-specific module; so it was necessary to develop one, along with a GZIP module to handle ARC GZ files, to have these features.

Apart from listing the properties to extract from the ARC files, the challenges that appeared at the design stages were mainly linked to the interpretation of the often ambiguous ARC specification.

First, we defined a unique, unambiguous terminology for the constituent parts of an ARC file:

– *Version-block*: the header and structure declaration of the file; comprises a *filedesc* (metadata about the ARC file) and a *URL-record-definition* (structure of the ARC records).

– *ARC record*: a specific entry of an ARC file, comprising a *URL record* (header for the ARC record) and a *network doc* (whatever the protocol returned to the crawler). This *network doc* is itself divided into a *protocol response* and an *object* (the harvested file).

We typically encountered difficulties in finding a way to manage the peculiarities of the ARC 1.1 format, an Internet Archive ARC profile with an XML metadata file inserted after the *filedesc*. Even though not referenced in the ARC specification, it was assumed to be compliant with it, since its author, IA, produces all its ARC files produced on this model since 2005. However, aligning this to our terminology was not simple: should this XML file be considered as a part of the *version block*, or as a particular *ARC record*? We combined the two, considering the *version-block* as a header built on the structure of an ARC record with an XML file as a possible content *object*, as can be seen on figure 2.

The other major problem was handling the gzip compression of an ARC: whereas a gzip compression is typically applied after the file has been created, Heritrix directly interlaces the two formats when creating arc.gz files: the version block and first ARC record

[5] NetarchiveSuite website: http://netarchive.dk/suite.

[6] An ARC file is a container file aggregating each file harvested on the Web in a dedicated ARC record. For technical reasons, the size of an ARC file is generally limited to 100 Mb. Cf. http://www.archive.org/web/researcher/ArcFileFormat.php.

[7] Jhove2 website: http://www.jhove2.org.

are respectively the first gzip and second gzip members at the same time. Jhove2 therefore had to be able to process an arc.gz file simultaneously with the gzip *and* ARC modules. We modelled this as an ARC structure with a gzip encoding:

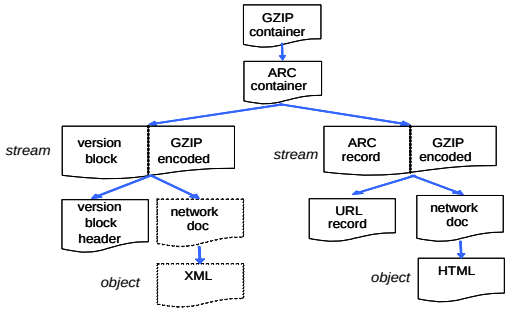


Figure 2. Structure of an arc.gz file according to Jhove2

2.2 Aligning the heterogeneous web archive collections: the NetarchiveSuite target

As explained in 1.1, BnF currently uses NetarchiveSuite, or NAS, for all its crawls, and the data harvested thanks to NAS represent the only growing part of BnF web collections. These are the two reasons why NetarchiveSuite data structure and data model have been chosen as a target to organize all our other collections of web archives in SPAR. For example, all the metadata describing Heritrix crawling process (configuration, log and report files from the crawler, or CRL) are packaged by NAS into ARC "metadata files", where each CRL file represents an ARC record within the ARC container file. We applied this rule to the CRL files coming from other harvesting channels whenever possible; we do not for example have any of them for the "historical" collections.

Another critical choice was related to the collections data model. The data coming from NAS are organized in three layers of granularity:

- At the base there is the ARC container file.
- Each ARC file is part of a specific "harvest instance", or "job". In our domain-specific terminology, a job is "a particular harvest process, realized by a crawling machine and monitored by a human engineer, which produces a set of data and metadata ARC files, and that has a beginning and an end". Each job is launched on a list of seeds (a seed is a URL from where the robot will start its crawling process), with defined parameters that will order the robot to conform to certain rules[8].
- On the top, the "harvest definition" is globally equivalent to a collection. A harvest definition groups all the jobs that have been launched in order to achieve the same purpose. For example, the aim of "performing a .fr domain crawl" is achieved thanks to hundred of jobs, each of them grouping thousands of domains. The harvest definition "news websites" launches every day a crawl of around 80 seeds – i.e. there are 365 jobs a year to achieve this harvest definition.

These three layers match three kinds of information packages in SPAR: ARC "data" files are ingested as "web data" information packages; ARC metadata files (that contain information at the job level) are ingested as "harvest metadata", and harvest definitions are ingested as OAIS Archival Information Collections.

In order to homogenize our collections to a certain extent, we decided to use this three layered data model everywhere, which can sometimes be artificial. For example historical collections only have two layers: the ARC files and the harvest definition. There is no layer for the job, as the data given to BnF has not been crawled, but extracted from a larger web archives collection. However, in order to maintain homogeneity, we virtually created a third layer. We declared that all the harvest definitions of the historical collections had been produced by a single harvest instance, or job.

2.3 From web archive concepts to PREMIS

Even if PREMIS is conceived as core preservation metadata and web archive-specific concepts are clearly out of its scope, it is a cornerstone of the SPAR data model [1] so we had to know where our web archive concepts fit in the PREMIS data model. Here again we encountered some difficulties.

The job. We have defined in 2.2 what a job is. However, if we try to fit this "job" concept in PREMIS, it can match three different entities depending on what you consider. It is a typical Event since it has a beginning and end date, produces Objects (ARC files) and has Agents (software, administrators…) involved in it. But it is also an Object, if we use this term to designate the result of a crawl. In addition, the job is also a set of parameters, given to a crawler at a certain time and impacting the crawling event and produced objects. From this standpoint, a job can be viewed as an Agent that activates a crawl.

These ambiguities forced us to adopt a clearer, PREMIS-compliant terminology:

– The Event is a **harvest** eventType.
– The Objects produced by this harvest event are **harvest instances**. They typically consist of at least one ARC metadata file and many ARC data files.
– The Agent activating a harvest Event launching a crawler is a **job**. Since it is currently impossible to track accurately – as the parameters can be changed by an administrator during a crawl – it was considered out of the scope of our digital repository, so we merely kept track of it as a linking Agent of the harvest Event. However we kept track of two key parameters: the user agent and the robots policy.

The user agent is an identity under which the crawler declares itself, typically a particular web browser, e.g. "Mozilla 5.0". We modelled this as a distinct Agent involved in the Event, because the same crawler could declare itself under different identities.

The robots policy is the behaviour of the crawler towards the robots.txt protocol (comply with it or ignore it). We considered it as a particular outcome of the harvest; this debatable choice was made in want of a current PREMIS field for "input" information.

The reports on the produced ARC files and crawled hosts were typical outcomes of the harvest Event, documented in Extensions.

[8] For example: do only crawl URLs in a given list of domain names, do not follow too many clicks from a seed URL…

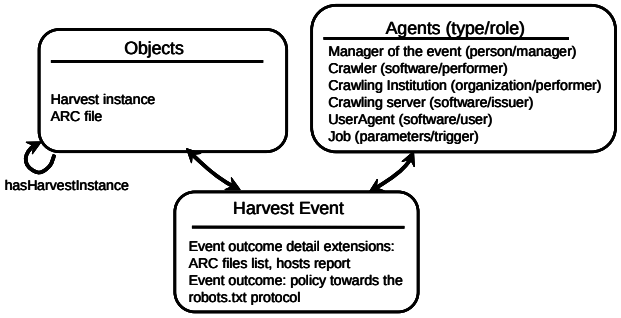


Figure 3. Aligning web archiving concepts with PREMIS

2.4 From core to domain-specific preservation metadata: the containerMD solution

Having validation and extraction tools for ARC and arc.gz files was not enough: we also had to define how to record this format specific information in our AIPs. While we had the XML output of Jhove2, it was considered far too verbose to be manageable[9].

Core information about a file and its content files could be modelled as premis:file containing premis:files[10]. However, some ARC-specific features needed dedicated fields to be recorded; and, above all, PREMIS has been designed on an all-or-none principle: it is necessary to choose between describing just the container file and describing all the contained files individually.

No characterization schema for container files being available to the community yet, we felt there was a gap to be bridged and we therefore designed containerMD[11]. Its key features are a description of the container file itself in a <container> section and two verbosity levels:

– A "non verbose mode": aggregated information about the entries in an <entriesInformation> section;

– A "verbose mode": dedicated description for each entry in an <entry> section, with the ability to include other characterization schemes if needed.

The <container>, <entriesInformation> and <entry> all have an extension section for format-specific fields, with only the ARC format for the moment. Each content *object* is thus referenced as an entry, with additional information about the ARC record being embedded in the ARC-specific extension section.

Figure 4. Aggregation methods in containerMD

	verbose mode: <entry>	**non verbose mode: <entriesInformation>**	**Reduction method**
existence of an entry	One <entry> per entry	number attribute	Count
creation dates	creationDateTime attribute	firstDateTime and lastDateTime attributes	Only the min and max values are kept
entry size	<fixity>: size attribute	minimumSize and maximumSize attributes	Only the min and max values are kept
format of a content file	<format>: name and version attributes	<formats> container element For each format name and (if any) version: one <format> element with name and version attributes; number attribute counting the corresponding entries; globalSize attribute for the size of all the corresponding entries.	Aggregation and count Sum
encodings at entry-level	<encoding>: type (encryption or compression) and method attributes	<encodings> container element For each encoding type and method, one <encoding> element with type and method attributes	Aggregation
ARC record host	<host>	<hosts> container element For each <host>, number attribute; globalSize attribute for the corresponding entries	Aggregation and count Sum
ARC record declared MIME type	<declaredMimeType>	<declaredMimeTypes> container element For each <declaredMimeType>, number attribute; globalSize attribute for the corresponding entries.	Aggregation and count Sum
ARC record protocol information	<response>: protocolName and protocolVersion attributes	<responses> container element For each <response> with a particular protocolName and (if any) protocolVersion: number attribute; globalSize attribute for the corresponding entries.	Aggregation and count Sum

3. CONCLUSION: FUTURE USAGES AND EVOLUTIONS

In the end, the ingest of the BnF web archives in SPAR will allow us to build **new curation services for the web harvesting team**:

- **Getting better file formats statistics** on the type of files (text, image, video…) harvested: currently we still use the MIME type sent by the server, which is often unreliable. Using Jhove2 and storing the results in the containerMD <formats> section to be queried will improve this knowledge.
- **Knowing the content of older collections**. The distribution of the data per host is also some key information for web archives. This information is compiled in files called hosts-reports for current harvest instances, but not for historical collections. Jhove2 will be able to calculate a host-report per ARC file, which may later be aggregated at upper levels.
- **Checking collection comprehensiveness**. 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 that we have lost data.

All this generated AIP metadata will also help us define indicators and **investigate preservation strategies**. Some metadata elements can be used to define risk assessment criteria, e.g. the format of the container file and its content objects, the rendering tool intended for the harvested files (given by the user agent), or the status of a given harvest (finished, terminated or crashed). This will help us define subsets of our collection on which focused preservation actions could be performed: format migration for the container files or not; emulation vs migration of the harvested files, and so on.

Finally, one of the great strengths of SPAR being its ability to manage different collections with different service level agreements, one may imagine **applying differentiated SLAs** to collections that, even though produced by the same harvesting infrastructure, do not share the same preservation policies. For

[9] Even with mere identification of the content files, the Jhove2 XML output for an ARC file could sometimes exceed the ARC file size itself. For the typical 100 Mb ARC file, this was considered too heavy to handle (processing, rendering, etc.).

[10] Cf. *Data dictionary for Preservation Metadata: PREMIS version 2.1*, p. 45. Online: http://www.loc.gov/standards/premis/v2/premis-dd-2-1.pdf.

[11] For more information on containerMD, see http://bibnum.bnf.fr/containerMD.

example, if we negotiate with a publisher to harvest PDF online editions provided that they comply with a specific PDF profile, we will be able to ask SPAR for stronger validation procedures to check that these files conform to the negotiated format. In the end, defining differentiated preservation actions and services on our web archives seems to be the next great challenge to take up.

4. REFERENCES

[1] Fauduet, L., Peyrard, S., A data-first preservation strategy: data management in SPAR, *Proceedings of the 7th International Conference on Preservation of Digital Objects (iPRES)*, 2010. Online: http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/fauduet-13.pdf (accessed September 2011, 30th).

[2] Illien, G., Sanz, P., Sepetjan, S., Stirling, P., The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future, *Proceedings of the 77th IFLA General Conference and Assembly*, 2011. Online: http://conference.ifla.org/sites/default/files/files/papers/ifla77/193-stirling-en.pdf (accessed September 2011, 30th).

[3] Kimpton, M., Braggs, M. and Ubois, J. 2006. Year by Year: From an Archive of the Internet to an Archive on the Internet. In *Web Archiving*, J. Masanès, Ed. Springer, Berlin, Heidelberg, New York.

[4] Lasfargues, F., Oury, C., Wendland, B., Legal deposit of the French Web: harvesting strategies for a national domain, *Proceedings of the 8th* International Web Archiving Workshop (IWAW), 2008. Online: http://iwaw.net/08/IWAW2008-Lasfargues.pdf (Accessed September 2011, 30th).

[5] Oury, C., « Large-scale collections under the magnifying glass: format identification for web archives », *Proceedings of the 7th International Conference on Preservation of Digital Objects (iPRES)*, 2010.

Virtual Archiving for Public Opinion Polls

Jonathan Crabtree
University of North Carolina
Odum Institute
22 Manning Hall
Chapel Hill NC USA
+1 919 428 6112
Jonathan_Crabtree@unc.edu

ABSTRACT

The Odum Institute for Research in Social Science Data Archive at the University of North Carolina, and partners from the National Network of State Polls present progress on a two year demonstration project using the Dataverse Virtual Archiving technology [1]. The goal of the Virtual Archiving for Public Opinion Polls: A Demonstration Project aims to streamline the ingest process and increase timely submission to data archives. Bridging this gap between producers and archives will increase the overall submission rates and ultimately preserve many data sets that would otherwise be lost.

Around the world researchers and scientists collect vast amounts of data, which often are not archived after the completion of the project or task. As researchers move on to new projects, past data they have collected are seldom documented and preserved [6]. Until the tools for data curation are integrated into the research lifecycle of data, we will continue to experience this problem [2]. This project seeks to provide a solution for this problem. Although the virtual archiving technology needed to bridge the gap between the data producers and archives already exists, the availability of this tool and its value needs to be communicated to the scholarly community.

The technology we use for this demonstration can be applied to many disciplines and data types. In this demonstration, we use public opinion data producers because these data serve as a useful, readily recognized example that will be widely replicated. Public opinion survey data are the most prevalent single kind of social science data and usually what most scientists first encounter.

The Odum Institute's relationship with the various state polling agencies and the National Network of State Polls make it an ideal candidate to propose new and innovative changes in the data life cycle of public opinion polls. This projects builds on our previous work with the Dataverse Network (DVN), developed at Harvard University. The Odum Institute has been an active partner in the DVN development and has recommended system modifications to allow for the maximum flexibility in public opinion preservation, distribution, and analysis. The DVN provides the tools necessary to implement this change. In this project, we use the DVN technology to aid data producing agencies in the ingest, curation and preservation of public opinion data, election polls and reports.

Automated ingest tools specifically designed for quantitative data are used to create metadata automatically on ingest. This is critical for two reasons. First, metadata are essential in making data accessible to the scholarly community beyond those who were involved in the data collection. Second, the creation of metadata — itself a technical field with its own set of tools and norms — is a specialty that lies beyond the expertise of most research teams.

The project is creating customized Web interfaces or "virtual archives" for each of the participating data producers. These interfaces are created to allow the researchers to seamlessly upload their data. The data will be transparently archived, preserved and curated by an organization trained in the field of long-term preservation. Once in the Dataverse Network, the data can be discovered and accessed via the existing federated search capabilities of the DVN. We are designing new workflows and help train the participating data producers to use these new and efficient methods of data ingest.

Categories and Subject Descriptors

H.4.1 [**Office Automation**]: Workflow management

General Terms

Management, Documentation, Design, Human Factors, Standardization.

Keywords

Digital Archives, Alliances, Federation, Data Management, Social Science Data

iPRES2011, Nov. 1–4, 2011, Singapore.

1. INTRODUCTION

Researchers and scientists collect vast amounts of data worldwide. Often these data are not archived or preserved following completion of the primary task for which they were intended. A number of scholars in their respective disciplines are champions for the difficult tasks of documentation and archiving; however, these tasks seldom receive funding and are often the first items cut from budgets.

Though the list of explanations can be expansive, the research community will continue to experience this issue until the tools for curating data are integrated into the research lifecycle of data.

The virtual archiving technology required to bridge the gap between the data producers and archives already exists [3]; the necessary next step is enhanced awareness of the accessibility of this technology and communicating its value to both scholars and the academic community.

Data-producing organizations are typically supportive of archiving the materials produced. Economic and workflow issues tend to inhibit attempts at comprehensive archiving. In order to maintain the economic feasibility of their data collection organizations, researchers must often move from one project to the next with minimal downtime often undermining attempts to adequately archive valuable data. The costs associated with rehiring qualified staff when new projects arise and retaining staff on the payroll with no outside financial support are considerable. Organizations must consistently maintain a queue of new projects and opportunities in order to justify their continued existence; a strategy that reserves little time and resources for data archiving.

The collection of social science research data — and particularly public opinion data — is ensnared in this problematic process, resulting in the loss of numerous valuable datasets [5]. Access to empirical social science data is fundamental to successful social science policy development, research and education. For example, students and teachers who wish to gain a deeper understanding of the findings in economics, psychology, political science, sociology, educational research, and other social sciences must be able to discover and access the data that constitute these studies. Teachers and students in the natural sciences also routinely encounter the products of empirical social science in surveys, newspaper editorials, magazine articles and other academic research products. The data that underpin many social science research studies discoveries and theories have not been consistently archived despite mandates from funding agencies such as the National Institutes of Health and National Science Foundation. This is due primarily to the enormous degree of post-project effort required to prepare data for archiving. To help ease this burden, data archive managers and data producers must work cooperatively.

2. VIRTUAL ARCHIVING

The Virtual Archiving for Public Opinion Polls project demonstrates a streamlined process for data submission from polling agencies across the country. The Institute will develop virtual archives for data producers to facilitate simple, direct access to the submission process. In traditional research data archival workflows, materials arrive at the archives after the project is completed. Ideally, the research teams would assemble the data, materials and any existing documentation post-project. Then, the materials are forwarded to the archive for ingest processing. Ingesting involves preparing data for archiving, de-identifying personal and confidential information, creating standard file formats, building any necessary metadata and documenting this process. The depth and quality of the ingest process varies greatly in each situation, and the effort required to assemble the components often limits the amount of materials archived. In most cases, the researchers have already moved onto new projects and do not have the time to follow through with the archiving steps.

With virtual archiving, the researchers begin using the archival tools earlier in the research process. These simple Web-based tools allow researchers and their staff to manage their data and documentation throughout the life cycle of the project. The virtual archives that result recreate the look and feel of the home institution Web sites. Simple ingest procedures provide metadata validation routines that assist in documentation and even prompt researchers to enhance their metadata. When quantitative data is ingested, automated routines create detailed variable-level metadata without requiring costly manual procedures.

The goal of virtual archiving is to provide simple tools that research teams can use to easily manage their data. As these research teams begin submitting their datasets, the ingest tools collect and verify much of the required metadata. When the research team releases the dataset by setting appropriate permissions, the archival submission process is complete. Although the process seems to the research team to be local on their Web site, the data is stored in the remote archive site. The data is backed up and preserved in a trusted replicated network from the moment it is ingested until it is released to the public. Trained archivists manage the process and ensure that important documentation and archival formats are created to ensure proper preservation. After datasets are archived, users will be able to search for the data from the producer's local Web sites and other scientists will be able to discover these studies and harvest the metadata using the Dataverse Network. Credit and acknowledgement for the data will remain with the research teams who produced the original data. The virtual archives are part of a national federated network of social science archives that aid in the dissemination of the work.

The demonstration project is developing archival and ingest workflows for five social science polling centers: the University of South Carolina Institute for Public Service and Policy Research, Monmouth University Polling Institute, the University of Georgia Survey Research Center, the University of Arkansas and the University of Indiana Center for Survey Research. Once these demonstration sites are complete, the Odum Institute plans to implement this technology at other national state polling centers.

This project focuses on public opinion or election data and the polling agencies that collect them, but the virtual technology involved — as well as our open source and distributed acquisition method — can be applied to other disciplines and data producing organizations. Though many of the features and analysis components are geared toward quantitative data, the virtual archive process remains applicable to qualitative data, documents, and images. A virtual curated preservation environment will promote effective and timely archival dataset preservation. The ability to customize virtual archives to meet the needs of individual data producers adds to the value of this workflow system. The benefits of a simple ingest workflow for datasets is considerable. Breaking down the barriers to ingesting data will ensure that a significantly larger portion of research data are archived properly. Though the precise impact of this virtual ingest demonstration may be difficult to estimate, we anticipate that archive submissions will increase by more than fifty percent.

3. PROJECT DESIGN

The project is developing virtual archives of election and public opinion poll data, a versatile system that will assist data producers across multiple disciplines.

The primary project goals are:

- Demonstrate use of the Dataverse Network and virtual archives to streamline the submission workflow process. These virtual archives will provide data producers tools for ingest, automated metadata creation and validation using Web-based client-server technology while preserving the look and feel of their local Web site storage. These Web-based workflows will allow data producers the ability to upload their datasets in the archive and document them in a seamless client-server environment. Once the data is archived, producers can assign rights, analyze and manage their data using the many Web-based services offered by the DVN;
- Build generic virtual archive template models to allow simple adoption of DVN technology. Templates are a predetermined set of code generic enough to fit applications across repositories and domains. They will be built using Java code, XML and HTML and can be easily modified to accommodate colors, logos, headers and footers to readily create the look and feel of the home institution Web site. These templates will reduce the cost of future virtual archive creation and integration;
- Work with polling data producers in an effort to increase archival rates;
- Train data producers in the use of quantitative automated ingest tools; and
- Disseminate findings and experiences to the preservation and data producing communities.

In addition to the initial, in-depth evaluation processes, the project consists of four major areas of effort: research and design, programming, training and reporting. The work plan begins by evaluating producing agencies to ensure that programming and design take full advantage of similarities across sites. Evaluation is not necessarily a onetime, linear process; findings from later phases will inform the ongoing design and programming phases.

Phase I: Research and Design

Odum programming staff and research assistants are working in conjunction with, and seeking input from, individual data producing agencies to understand existing workflows and processes local to each demonstration site. The mission for this phase is to understand the local environments of the demonstration sites, design individual virtual archives for these sites and seek commonalities for use in designing templates to reduce the cost of future virtual archive design. Odum staff seeks to find similarities and efficiencies in the design of the virtual archives for the demonstration sites. These commonalities will be used to create base programming templates during the programming process. This phase will also involve collecting baseline quantitative information from each producing agency on how many datasets they have archived. This will allow us to compare the numbers of datasets archived (and how long archiving took) by each agency before and after implementing the virtual preservation process. This information will allow us to ultimately quantify our results and identify a metric of success.

Phase II: Programming

Odum archive staff are creating virtual archives within the Odum Dataverse Network archive software for each demonstration site. These archives are customized portions of our archival system designed to house the producer's data. Once a dataset is part of the system, data producers can define access controls, download data, analyze data using statistics and promote international discovery of their data though the Institute's federated archival network. Odum programmers work with designers to construct Web-based interfaces for the new virtual archives. These interfaces integrate the demonstration partners' existing Web sites and provide continuity of appearance and function for the researchers and users. The Institute will incorporate ongoing recommendations from the individual data producers to provide a streamlined ingest workflow and minimize barriers to submission. Developing these Web-based interfaces requires customized programming that can be costly and inhibitive to widespread adoption of the technology. For this reason, we will create programming templates that build on the commonalities among the participants and will use these similarities to create code templates reducing the future cost of virtual archive integration. Templates will reduce costs and will provide a base for future dissemination of the technology to a wider community. We are document the programming process to assist in developing training materials in the next phase.

Phase III: Training

Institute staff are designing training documents and visual aids for data producers and will help train remotely the data producers at the demonstration sites in using the automated ingest tools and metadata templates. In addition to hands on training, Web-based video instruction will be used to provide economical and effective training. The Institute will produce and disseminate live Internet streaming of Institute short courses designed to educate data producers on the virtual archive workflow process. Once training is complete, project staff will supervise and provide ongoing technical support for the demonstration sites.

Phase IV: Reporting

This final phase will assemble reports for the sponsor as well as develop and execute the final evaluation surveys. Project staff will work with the Odum Institute survey design and methodology staff to develop these surveys, which will gather information on how many datasets, are being archived and how long the process takes for each data producer. Following this phase, the Odum staff and data producers will leverage the existing social network within the National Network of State Polls to help disseminate the findings both nationally and internationally to archivists and data producers.

4. CURRENT PROGRESS

The project is nearing the end of the first year and has been very successful to date. Qualitative interviews with each polling agency have been conducted and provided to the design team with information to begin the programming process. Programming has been completed for four of the NNSP partners with one of the virtual archives already in use. The team has designed the Web interface for the individual virtual archive and developed ingest templates for the different survey methodologies used by the agency. These templates record commonly used metadata responses to enable polling agencies to streamline the ingest process and easily train their staff to use the virtual archive for ongoing data management and documentation by simplifying the process. Initial reactions to the interface and processes have been very favorable.

Current efforts include finalizing programming on the remaining virtual archive interfaces and developing training documentation using the feedback from the initial deployments. We are developing training videos for web-streaming application to assist in the education of our participating agencies and for continued use in the dissemination of this technology.

5. CONCLUSION

The research community is faced with expanding burgeoning collections of digital data. As researchers and scientists struggle to deal with this vast amount of information, they still have to continue their primary scientific work and would like assistance in this process [7]. A recent poll of *Science's* peer reviewers shows that 20% of those asked were creating data sets larger than 100 gigabytes and 7% used data sets greater than 1 terabyte [7]. When asked where the respondents archive the data created by their research, over 50% claimed they stored the data in their labs [7]. Additionally, 38.5% reported that they archived their data on university servers while only 7.6% used community repositories [7]. This leaves most data to reside outside of archival repositories and beyond the care of data curators. This lack of stewardship places much data at risk and raises the questions of "what roles can digital archives play in the preservation process and when should they become involved in the data lifecycle". This project seeks to insert archival processes earlier in the research data lifecycle in the hopes of ingesting a larger portion of these valuable projects. Early indications are very positive and data producers are very open to examining this change in workflow. Current demands by funding agencies for research data management plans have forced researchers to think about the preservation of their data during the proposal development process [4]. Archives should provide assistance in this process and need to provide tools that aid in reducing the efforts require managing these data. Virtual archives are a potential solution for many researchers. This project seeks to demonstrate the usefulness of this technology and document the workflow process. Early responses are very favorable and we have been approach by additional agencies wishing to examine the tools.

6. ACKNOWLEDGMENTS

I would like to convey a world of gratitude to the Odum Institute for Research in Social Science for the freedom to work on projects advancing digital archival research. I would also like to thank all our Data-PASS partners for assistance in the common goal of better social science data preservation and access. In addition, I would like to thank the Institute for Museums and Library Services for the funding of the deployment of virtual archives in a demonstration project for public opinion data collection centers. IMLS National Leadership Grant 2010: Award Number LG-07-10-0240-10

7. REFERENCES

[1] DVN, Dataverse Network Repository Software,www.thedata.org

[2] Green, A.G. and Gutmann, M. (2007). "Building Partnerships Among Social Science Researchers, Institution-based Repositories and Domain Specific Data Archives." *OCLC Systems & Services: International Digital Library Perspectives* 23:35-53.

[3] King, G., (2007), An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. *Sociological Methods & Research*, *36*(2), 173-199. Retrieved March 21, 2008, from http://gking.harvard.edu/files/dvn.pdf

[4] NSF, National Science Foundation Data Management Plan requirements, Retrieved March 2011 from http://www.nsf.gov/bfa/dias/policy/dmp.jsp

[5] Parry, J., Kisida, B., and Langley, R. (2006). "What Would (Mac) Jewell Do? The State of State Polls" Presented at the Annual Meeting of the Southern Political Science Association Atlanta, Georgia January 2006. Accessed January 24, 2010 from http://www.allacademic.com/meta/p_mla_apa_research_citation/0/6/8/7/7/pages68776/p68776-1.php

[6] Pienta, A. M., Gutmann, M. P., Hoelter, L. F., Lyle, J. and Donakowski, D. (2008). "The LEADS Database at ICPSR: Identifying Important "At Risk" Social Science Data." Paper presented at the annual meeting of the American Sociological Association Annual Meeting, Sheraton Boston and the Boston Marriott Copley Place, Boston, MA. Accessed January 15, 2009 from http://www.allacademic.com/meta/p242699_index.html.

[7] Science (2011), Challenges and Opportunities, *Science 11* February 2011: Vol. 331 no. 6018 pp. 692-693 DOI: 10.1126/science.331.6018.692, Retrieved March 26, 2011 from http://www.sciencemag.org/content/331/6018/692.short

Emulation Reading Room Prototype

Sebastian Schmelzer
Albert-Ludwigs University
Freiburg, Germany
sebastian.schmelzer@rz.uni-freiburg.de

Dirk von Suchodoletz
Albert-Ludwigs University
Freiburg, Germany
dirk.von.suchodoletz@rz.uni-freiburg.de

Klaus Rechert
Albert-Ludwigs University
Freiburg, Germany
klaus.rechert@rz.uni-freiburg.de

ABSTRACT

The electronic collections of today's libraries, museums and archives are growing and increasingly have a more relevant role in the holdings. Memory institutions must address users' need to access a widening range of digital artefacts. Often the formats of those artefacts are outdated and they cannot be run or rendered on today's systems any longer. This is where emulation can provide the required digital environments suitable for a given object type. Practical research is being done at Freiburg University for the Open Planets Foundation on how to integrate different emulators for a number of original environments into a single graphical desktop. In this case study, options for future reading room systems like stateless Linux workstations are evaluated and prototypical implementations are demonstrated.

1. INTRODUCTION

For modern memory institutions it would be desirable to offer their users access to a wider range of different object types within their original environment than currently possible in today's reading room systems, which are restricted to a number of multimedia formats and document types. Many artefacts of the growing electronic collections are outdated, meaning no modern application can render them in a usable and authentic way. These artefacts require appropriate front-ends in order to experience (render) them on today's systems.

The emulation of outdated digital ecosystems is the answer for accessing and experiencing an increasing variety of (otherwise inaccessible) file types. Compared to traditional file loading or program starting, the access to ancient artefacts involves a more complex workflow. On average, the visitor to memory institutions is not familiar with past computer architectures. The goal of researching the different approaches is to create different ways to automate access to the desired digital artefact. The concept is to have different prototypes for running such a service, such as "mimicking" the double-click on an object and starting an "automagic" loading into the proper application. This automation will assist the average user of a reading room digital object access machine, who is not very familiar with past GUI concepts and applications. Instead of simply loading a certain file into an appropriate application, a more complex procedure is to be executed in order to wrap the object and transport it into the original environment. Then the environment has to be turned on and the object loaded into it's original creating or viewing application.

iPRES2011, Nov. 1–4, 2011, Singapore.

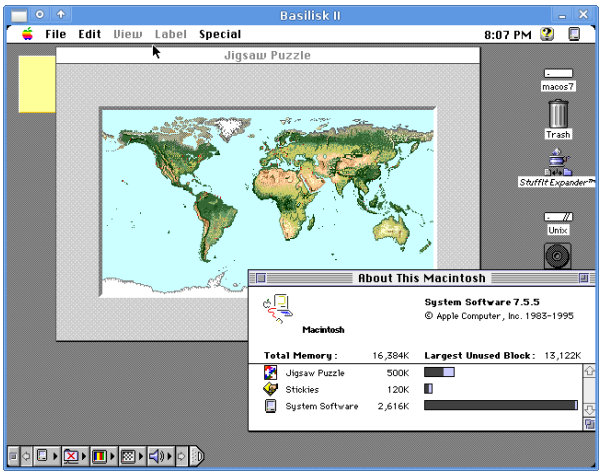


Figure 1: Mac OS 7 was a popular operating system used in many companies and institutions as well as by private individuals.

2. ACCESS TO DIFFERENT EMULATORS

There is a large number of emulators, mostly programmed by enthusiasts available as open source. Often, more than one emulator could be used for a certain digital ecosystem of hardware and software. Especially for the x86 architecture, there exists a wide range of commercial and free emulators and virtualization tools like VMware, VirtualBox, QEMU, Dioscuri or DosBox. To offer the user easy access to the different combinations of original environments and emulated hardware, we produced a small application (Fig. 2) that reads the metadata from an XML file in order to provide a short description plus the machine and firmware information needed for the original environment to start. Thus it is easily possible for the user to compare the characteristics of the same original environment in different virtual machines or evaluate the rendering or execution of a wide range of digital

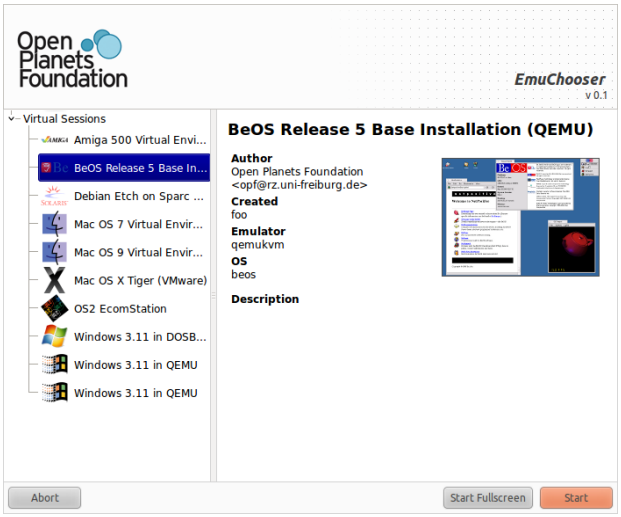


Figure 2: Selection to choose different original environments

artefacts [2]. The application is executed on the Linux platform, which is a very versatile base for hosting a wide range of emulators without incurring additional licensing costs in order to run a larger number of reading room workstations. Providing access to a wide range of different outdated computer platforms is the first part of an access strategy. As many users are no longer familiar with outdated architectures and user interfaces, the idea is to prepare the original environment, configure the emulator and load the artefact into the started environment automatically. After loading, the user information is provided with information on how s/he can use the artefact or navigate the original environment.

3. ASSISTED CREATE VIEW

Based on the concept of view paths, which formalize the pathway from a certain digital object like an Amiga computer game or Word Perfect document into its execution or rendering environment, we explore methods as to how this could be (partly) automated. A multitude of methods are being researched, ranging from altering the original environments in different ways to abstract automated handling of user interaction as discussed in [3]. Thus, ongoing research is looking into the application of automation procedures provided by the original operating systems and applications. This alters the approach chosen by Brown and Woods [5], who introduce custom binary executables and *AutoIt* scripts – a Windows GUI scripting package – into the original environments to automate the access to certain types of objects. Another option is to use VNC interfaces provided by the emulators or virtualization tools to run interactive environments unattended. This mechanism could be deployed to prepare the original environment in a way that avoids the need of any specific operating system or application knowledge. The system is developed in an abstract way in order to handle arbitrary environments and emulators. The concept of automated user interaction is broadened by looking into ways of providing the emulators with appropriate interfaces [1] or to explore the options of workflow automation more deeply by using the monitor interfaces provided by a number of emulators. These interfaces help to operate many of the emulator functions like (un)mounting removable media or sending keystrokes or mouse events in order to run certain actions unattended.

Many useful tools and applications are provided with a standard Linux installation: The preparation of floppy images or a wide range of container images with different filesystems is not a problem. Most of the relevant filesystems are directly supported by the Linux kernel which makes import and export of artefacts into and from the original environments easy. Additionally, converter tools like "qemu-img" or a wide range of decompressors help to program the several workflows involved. For the prototypical implementation we use the standard Linux scripting tools, or we programmed a small application such as the emulator chooser (Fig. 2). This tool provides the appropriate command lines pointing to the relevant resources like original system images and firmware roms or generates the configuration files needed for a range of virtualization tools. It could easily be extended to acquire license information from a centrally managed system. All sessions are run in non-persistent mode so that users can freely interact with the original systems without ruining the installations.

4. CONCLUSION

The direct access to many outdated environments still offers the most authentic experience. When provided within the premises of the memory institution, an emulation access workstation might be more favorable compared to remote access. It allows a more real-time, full screen experience with audio output and direct access to the peripherals if required. A stateless Linux system is an easy way to provide a reliable service enabling the users to fully interact with original environments run using today's hardware and does not risk rendering them unusable. The research described here does not focus on any file type detection during the whole procedure, but presumes that this information is available and encoded using the metadata of the files loaded. Besides a proper file format base and tool registry, a software archive of outdated software components is required [4].

5. REFERENCES

[1] Evgeni Genev. VNC Interface for Java X86-Emulator Dioscuri. Online, http://hdl.handle.net/10760/15102, October 2010.

[2] Mark Guttenbrunner, Christoph Becker, and Andreas Rauber. Keeping the game alive: Evaluating strategies for the preservation of console video games. *International Journal of Digital Curation*, 5(1), 2010.

[3] Klaus Rechert, Dirk von Suchodoletz, and Randolph Welte. Emulation based services in digital preservation. In *JCDL '10: Proceedings of the 10th annual joint conference on Digital libraries*, pages 365–368, New York, NY, USA, 2010. ACM.

[4] Maurice van den Dobbelsteen, Dirk von Suchodoletz, and Klaus Rechert. Software archives as a vital base for digital preservation strategies. Online, *http://hdl.handle.net/10760/14732*, July 2010.

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

Demo: Migration-by-Emulation

Isgandar Valizada
Albert-Ludwigs University
Freiburg, Germany
isgandar.valizada@rz.uni-freiburg.de

Klaus Rechert
Albert-Ludwigs University
Freiburg, Germany
klaus.rechert@rz.uni-freiburg.de

Dirk von Suchodoletz
Albert-Ludwigs University
Freiburg, Germany
dirk.von.suchodoletz@rz.uni-freiburg.de

ABSTRACT

The availability of migration tools for older formats is often limited. Thus we suggest a different approach: using the original applications to access the object and transfer the latter into formats which can be accessed in today's environments. The appropriate environment for the digital artefacts could be provided through emulation. With the reproduction of the original environment, a large and diverse set of migration input/output paths becomes available. Working for the Open Planets Project the authors the authors created remotely accessible Web services integrated into the PLANETS testbed. These services demonstrate preservation workflows using migration together with the emulation of original environments.

1. CONCEPT

A strategy for accessing digital artifacts with outdated formats is to convert them into formats which can be rendered or executed in todays digital ecosystems. In most cases the applications or operating systems developed by the software producers are the best candidates for handling them. Emulation is the best way to reproduce original digital environments, which themselves provide the base layer for very flexible multiple migration input/output scenarios. Typically, applications used to produce or render digital objects were programmed with a human user in mind. He operated the application through keyboard or mouse interaction. Many of those applications don't provide programming interfaces for unattended command line operation e.g. to perform a format migration. However, performing migrations manually for every digital object is not a feasible strategy in many cases; because of the large quantities held by many institutions, it may turn out to be a time-consuming and costly task. Additionally, depending on the original environments, many archivists or private users don't have the requisite knowledge on either how to install a certain application or operating system or how to handle a certain emulator.

iPRES2011, Nov. 1–4, 2011, Singapore.

The idea presented in this demo is the creation of easy-to-use migration tools out of a combination of original environments running in emulators and steered by automated UI interaction. In [3] the authors showed the general feasibility of recording and replaying interactive user sessions in an abstract way. We brought the development further by implementing services which conform to the PLANETS interoperability framework *migrate* Web Service interface (cf. [1]). Using this framework complex procedures to transform a digital object into a selected output format [2] could be deployed. In contrast to simple command-line input-output migration tools, a migration-by-emulation service needs a more complex initial setup:

- *System Emulation*: Hardware emulation including a full reconstruction of an ancient environment. For instance a i386 CPU, ISA Systembus, VESA compatible graphics, PS/2 mouse and AT keyboard are minimal requirements, e.g. for Windows 3.11.
- *System Environment*: An appropriate runtime environment (e.g. a disk image file) preconfigured with the operating system, necessary drivers and tools, and the required target application. Furthermore, each environment specifies at least one transportation option, defining how digital objects can be injected into and extracted from the virtual environment. Examples range from different kinds of floppy-disk images to hard-disk container formats and advanced networking options.
- *Interactive Workflow Description*: An abstract description of all interactive commands to be carried out in order to perform a certain migration. Such an description consists of an ordered list of interactive input actions (e.g. key strokes, mouse movements) and expected observable output from the environment (e.g. screen- or system-state) for synchronization purposes.

The created service is split into two parts. The scenario preparation unit (Fig. 1), typically run once, interactively records the user interaction and creates the event list for playback. The playback unit is used by the migration service and re-runs the once recorded events unattended.

2. MIGRATION-BY-EMULATION

As migration-by-emulation services should be accesible the same way as standard command line tools, they are registered and deployed using the same methods within the

PLANETS testbed [4]. The Java-programmed prototype for complex emulation-backed migration workflows has all core components implemented. The services are called from within the testbed standard procedures. Preconfigured original environments are deployed and the Grate-R VNC record service (Fig. 1) is used to generate abstract workflow descriptions. The generated interaction tracefiles were then attached manually to the appropriate original environment to form a migration unit. Two different migration services

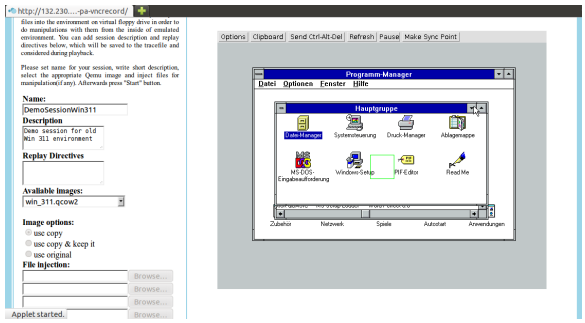


Figure 1: Scenario preparation unit and recording Web front-end

are registered within the testbed. One of them is a truly atomic migration accepting WPD as input and producing RTF as output. Thus the resulting file is directly delivered to the user after the procedure finishes. A migration took less then a minute per item and succeeded on a range of different input files. The second service is more complex as it takes an AMI Pro text document (SAM) as input and produces two different outputs, a TXT and a PDF (Fig. 2). The TXT is the result of a classical "save-as" migration. The PDF is generated by sending the document to a virtual printer generating PS as output. This file is then loaded onto the Ghostview application, which renders a PDF from it. This migration unit was deliberately of a more complex nature. We wanted to demonstrate the feasibility of producing more than one output file from a single input. This helps to evaluate and compare different workflows in regard to runtime, reliability and complexity. Nevertheless, the framework interfaces are to be extended to accommodate more flexible workflows which produce more than one result from a single input.

Additionally, a virtual disk-handling service was programmed to produce disk image containers for different emulators with the option to specify a range of supported filesystems understood by the original system environments. The creation of a QEMU compatible container with a FAT filesystem in it is comparatively simple; other containers and filesystems are supported to by using the "qemu-img" container conversion tool.

3. CONCLUSION

Our implementation focused on the feasibility of the preservation framework integration. Future research is dedicated to the speeding up of workflows by looking into the VNC recording and playback. The tracefiles are a good starting point for optimizaton. They could be enriched with additional metadata to use them for progress reporting. A certain state in the metadata directly correspondends to a state

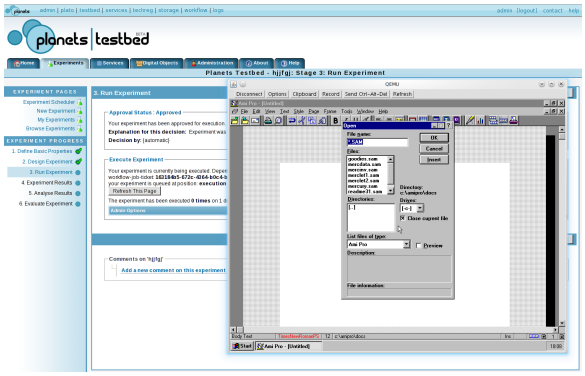


Figure 2: Migration experiment run in Planets testbed. QEMU interface connection opened for control and demonstration purposes.

of the migration workflow and could be reported back to the preservation framework. The tracefiles could be modularized to better identify the different stages, like original operating system booting, application starting, artefact loading, and saving in a new format. This information could be used not only for feedback but also to identify checkpoints. Those checkpoints could help with error recovery for restarting the procedure after failed attempts. Plus, these workflows could help to evaluate future versions of emulators before they get integrated into preservation systems. These issues are part of the ongoing research at Freiburg University.

With the integration of migration-by-emulation into the PLANETS testbed, such migration tasks become available to a wider community and hopefully encourages a range of different institutions to test, create and deploy such services. Since the individual migration services only require the preparation of system environments and the production of appropriate recordings, the proposed system is able to speed up the creation of diverse migration services, since no additional programming or integration effort is required.

4. REFERENCES

[1] Ross King, Rainer Schmidt, Andrew N. Jackson, Carl Wilson, and Fabian Steeg. The planets interoperability framework. In *Proceedings of the 13th European Conference on Digital Libraries (ECDL09)*, pages 425–428, 2009.

[2] Klaus Rechert, Dirk von Suchodoletz, and Randolph Welte. Emulation based services in digital preservation. In *JCDL '10: Proceedings of the 10th annual joint conference on Digital libraries*, pages 365–368, New York, NY, USA, 2010. ACM.

[3] Klaus Rechert, Dirk von Suchodoletz, Randolph Welte, Maurice van den Dobbelsteen, Bill Roberts, Jeffrey van der Hoeven, and Jasper Schroder. Novel workflows for abstract handling of complex interaction processes in digital preservation. In *Proceedings of the Sixth International Conference on Preservation of Digital Objects (iPRES09)*, 2009.

[4] Rainer Schmidt, Ross King, Andrew Jackson, Carl Wilson, Fabian Steeg, and Peter Melms. A Framework for Distributed Preservation Workflows. *International Journal of Digital Curation*, 5(1), 2010.

Re-awakening the Philips Videopac: From an old tape to a vintage feeling on a modern screen

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

Preserving data for a specific system usually depends on the system in question. Different strategies for different file types are necessary to preserve all data for a system. In this demonstration we present tools developed for preserving data for a home computer system from 1984. We present how a tool we developed can be used to retrieve data stored on audio tapes and how this data is migrated to non-obsolete formats. We also present how the data is migrated on a bit-stream level only and can then be used in an emulated environment using a recently developed emulator for the system's hardware. We further show the features of the emulator that allow its proper usage for digital preservation purposes. The purpose of the demo is to demonstrate different types of digital objects for a system, the information layers of these digital objects and how the proper preservation strategy is chosen. Preserving static digital objects and understanding the difference to preserve dynamic and interactive digital objects like software is a key preparation for the preservation of distributed software as in Software as a Service (SaaS) and even whole business processes.

1. INTRODUCTION

Digital objects can be separated in two different groups. One group contains static objects that are rendered by a viewer application. These objects can usually be migrated into a different format, so a different (non-obsolete) viewer application can be used to render them. The other group contains objects which either can't be easily migrated to a different format, as their behavior is also significant, or they are actually programs that need to be preserved. The typical preservation strategy for these objects is emulation. On any given computer system usually both types of objects exist: Data in the form of images, text-documents, spreadsheet data or database entries on one side and on the other hand digital objects like enterprise software applications, process management applications, interactive digital art and video games.

Figure 1: Philips Videopac+ G7400 with plugged in Philips C7420 Home Computer cartridge.

Concentrating on a comparatively simple system, a home computer system from 1984, we take a closer look at the different types of digital objects that exist for this system. We demonstrate two different applications that we developed in the past and that show two different preservation strategies: a migration tool that converts data stored by the original system to non-obsolete formats, and an emulator that allows us to execute original software in an emulated environment, but also allows us to manipulate and render data for the system using the applications the data has been created with. We show the requirements that an emulated environment has to fulfill to be properly usable for digital preservation purposes.

This demonstration proposal is structured as follows. First we present the home computer system and give an overview of the various data types that existed for the system. Next we present the two different tools for migration and emulation. Finally we discuss what we will have shown in the demonstration and how it aids the participants in their work.

2. THE HOME COMPUTER SYSTEM AND ITS DATA FORMATS

For our case study and demonstration we concentrate on an early home computer system manufactured by Philips in 1984. The Philips G7400 Videopac+ system was marketed as a video game system with a full keyboard that could be expanded to a full home computer system using the expansion module C7420 Home Computer cartridge as shown in Figure 1. The system itself was powered by an Intel 8048h processor, while the home computer cartridge used a Zilog Z80 microprocessor as it's main processing unit. For stor-

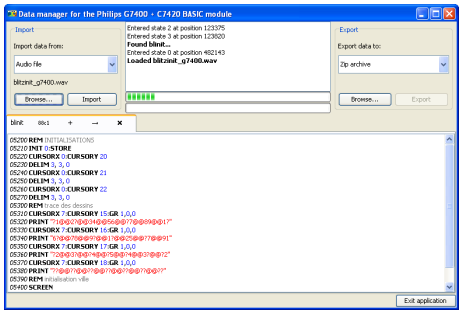


Figure 2: Migration tool for home computer data.

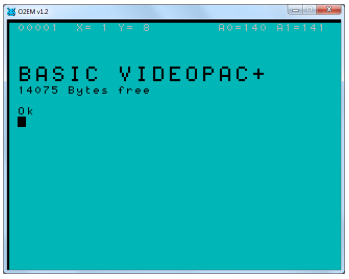


Figure 3: Start screen of C7420 Home Computer cartridge on O2EM emulator.

ing and loading data connectors to an external audio system (usually tape system) were provided.

The system was able to store the following different data formats as files on tape using different variations of its command for saving data 'CSAVE': Images (screenshots of the current screen), Arrays of data, Strings (up to 255 characters), Memory Dumps for raw binary data and BASIC programs to save any programs written in the system's programming language 'Microsoft BASIC-80'.

A detailed description of the available data formats and their storage in bit-format as well as the transformation between the analog audio signal used to store the data and the digital counterpart can be found in [2].

3. MIGRATING DATA TO NON-OBSOLETE FORMATS

The various data formats described in Section 2 most can be converted to a non-obsolete format. Even BASIC programs that are not directly runnable on a modern platform can be converted to text format. A migration tool that was presented in 2009 on iPres [1] and later in a paper for IJDC [2] was developed. It allows the migration of data between the original stored form in audio waves, the native bit-stream formats and non-obsolete formats (e.g. JPG for image data, text files for BASIC programs). A Screenshot of the migration tool is shown in Figure 2.

In the demonstration we will explain the difference between the formats (physical layer, logical (bit-stream) layer and conceptual format as well as discuss the context for the digital objects shown. The participants will learn to understand the difference and get to know which objects can be migrated and which can not.

4. RENDERING DATA IN AN EMULATED ENVIRONMENT

Some of the data objects retrieved from the storage medium are BASIC programs. Even though we will show how these can be migrated to human readable form in text format, they are not runnable on a current computer system. We show a recently developed emulator and demonstrate, how these programs can be executed in an emulated environment.

We will show some of the features present in the emulator that make it usable for digital preservation: Data injection using emulation of the original storage media, keyboard input and also the possibility to copy data from the host system into the emulated environment. Furthermore we show how data can be extracted from the emulated environment by copying the screen content to the clipboard, not only as screenshot but also in text format for further use. We explain how data that is stored in it's native bit-stream format and not migrated to a non-obsolete format can be rendered in the emulated environment. We also show how data that looses its context when migrated without the application creating the data can be loaded using the original (emulated) application and shown as interpreted by the original application.

5. CONCLUSIONS

The proposed demo shows participants the different data formats that exist for a computer system on the example of an early and comparatively simple home computer system from 1984. After the demonstration participants will understand the different layers of the data for the system. It will become clear how some digital objects can be migrated to non-obsolete formats while others have to be opened in an emulated environment. The usage of the emulator will show the complexity of emulation systems and the necessary features of emulators to support the work of archivists, librarians and any other party having to work with emulators in the digital preservation context. In the end of the demonstration we will discuss how the shown tools and methods translate to more complex systems to enable the preservation of standalone software but also for distributed software and enterprise application systems.

6. ACKNOWLEDGMENTS

The research was co-funded by COMET K1, FFG - Austrian Research Promotion Agency and by European Community under the IST Programme of the 7th FP for RTD - Project ICT-269940/TIMBUS.

7. REFERENCES

[1] M. Guttenbrunner, M. Ghete, A. John, C. Lederer, and A. Rauber. Digital archeology: Recovering digital objects from audio waveforms. In *Proceedings of the Sixth international Conference on Preservation of Digital Objects (iPRES 2009)*, pages 90–97, San Francisco, USA, October 2009.

[2] M. Guttenbrunner, M. Ghete, A. John, C. Lederer, and A. Rauber. Migrating home computer audio waveforms to digital objects: A case study on digital archaeology. *International Journal of Digital Curation (IJDC)*, 6(1):79–98, 2011.

Meet RODA, a Full-Fledged Digital Repository for Long-Term Preservation

Rui Castro
KEEP SOLUTIONS
Rua Rosalvo de Almeida, nº 5
4710-029 Braga, Portugal
+351 253066735
rcastro@keep.pt

Luís Faria
KEEP SOLUTIONS
Rua Rosalvo de Almeida, nº5
4710-029 Braga, Portugal
+351 253066735
lfaria@keep.pt

Miguel Ferreira
KEEP SOLUTIONS
Rua Rosalvo de Almeida, nº5
4710-029 Braga, Portugal
+351 253066735
mferreira@keep.pt

ABSTRACT
RODA is an open-source full-fledged digital preservation repository capable of ingesting, managing and providing continuous access to various types of digital objects, namely text-documents, raster images, relational databases, video and audio.

It is supported by open-source technologies and makes use of existing standards such as the OAIS [1], METS [2], EAD [3] and PREMIS [4].

Categories and Subject Descriptors
H.3.7 [**INFORMATION STORAGE AND RETRIEVAL**]: Digital Libraries – *Collection, Dissemination, Standards, System issues, User issues.*

General Terms
Management, Standardization, Design, Security.

Keywords
Digital preservation, authenticity, digital archive, digital objects, open-source, digital repository.

1. Introduction
RODA is an open source digital repository specially designed for archives, with long-term preservation and authenticity as its primary objectives. Created by the Portuguese National Archives in partnership with the University of Minho, it was designed to support the most recent archival standards and become a trustworthy digital repository.

RODA was developed on top of Fedora Commons and embodies high-level standards of security, scalability and usability. Its centralized architecture enables an easy management while the auto-deposit tools and ingest workflow account for the scalability of the human-resources.

RODA is a complete digital repository system that provides functionality for all of the main units that compose the OAIS reference model. RODA fully implements an Ingest workflow that validates SIPs and migrates digital objects to preservation formats, and provides Access by delivering different ways to search and navigate over available data as well as visualizing and downloading stored digital material.

Data Management functionalities allow archivists to create and modify descriptive metadata and define rules for preservation interventions, e.g. scheduling integrity checks on stored digital objects or initiate a migration process.

Administration procedures allow the definition of access rights to data and operational permissions for each user or group.

In this demonstration we will explore all the functionality provided by RODA: ingest, access, description, management and preservation by means of its well designed graphical user interfaces (Figure 1).

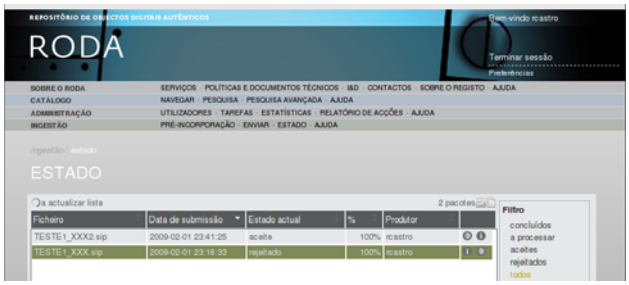


Figure 1 - Screenshot of RODA Web user interface.

2. Features to be demonstrated
In this section we describe the features of the system that will be demonstrated.

2.1 Ingest workflow
New entries to the repository come in the shape of Submission Information Packages (SIP). When the ingest process terminates, SIPs are transformed into Archival Information Packages (AIP), i.e. the actual packages that will be kept in the repository. Associated with the AIP is the structural, technical and preservation metadata, as they are essential for carrying out preservation activities.

The SIP is composed of one or more digital representations and all of the associated metadata, packaged inside a METS envelope. Producers take advantage of a small application called *RODA-in* that allows them to create these packages. The SIP is a

iPRES2011, Nov. 1–4, 2011, Singapore.

compressed ZIP file containing a METS envelope, the set of files that compose the representations and a series of metadata records. Within the SIP there should be at least one descriptive metadata record in EAD-Component format.

Before SIPs can be fully incorporated into the repository they are submitted to a series of tests to assess its integrity, completeness and conformity to the ingest policy. After decompressing the SIP, the validation process performs the following set of tasks: virus check, envelope syntax check, SIP completeness check, file integrity check, descriptive metadata check, preservation metadata check, representation check, specific representation check and format normalization. If all of these steps are completed successfully, the SIP waits for acceptance, which can be done manually or automatically (in batch mode).

After a successful validation, the SIP is then taken apart and each of its constituents is integrated into the repository. After this procedure, the SIP becomes an AIP and is ready to be disseminated to potential consumers that have clearance to access that information.

2.2 Access

Consumers are able to browse over available collections to view or download digital representations stored in the repository.

Depending on the type of the digital object, different viewers/disseminators are used. For example, text documents are delivered to consumers using a flash-based page-flip Web application and as a download so the consumer can use its favourite PDF viewing application. Representations composed of several images (e.g. digitised works) are displayed in a special Web viewing application that allows consumers to navigate through the pages of the representation.

2.3 Data management and archival storage

Because RODA was developed to be used by the Archival community, the underlying descriptive metadata supported by the repository is EAD/XML (Encoded Archival Description).

RODA's content model is atomistic and very much PREMIS-oriented. Each intellectual entity is described by an EAD-component. These metadata records are organized hierarchically in order to constitute a full archival description, but are kept separately within the Fedora Commons content model.

PREMIS entities are used intensively as they shape the overall content model of the repository. Archival storage is supported entirely by Fedora Commons.

2.4 Administration and Preservation planning

Administration allows for fine-grained control of user access to information and functionalities. Permissions can be granted to users or groups to perform certain actions within the repository and to access certain data or metadata objects.

All actions performed in the repository by any user are logged *ad aeternum* for security and authenticity reasons. Every user must be authenticated in order to use the repository.

Preservation management within RODA is handled by scheduled events. A preservation expert defines the set of rules that trigger specific preservation actions and when these should take place. Preservation actions comply to a common API, so creating and installing new preservation actions in the repository is as easy as copying the programme file to the correct directory on the server. Preservation actions include format migrators, checksum verification tools, reporting tools (e.g. to automatically send SIP acceptance/rejection emails), etc.

Each preservation event that takes place inside the repository is recorded as preservation metadata. Special events like format migrations establish relationships between representation nodes and record agents used and event outcomes automatically.

3. Architecture

RODA's architecture is service oriented and it's composed of 3 main components: RODA WUI, RODA Core Services, and RODA Migration Services.

RODA's Web User Interface (RODA-WUI) handles all the aspects of the graphic user interface for producers, consumers, archivists, system administrators and preservation experts. The RODA-WUI components are supported by AJAX and Web services. RODA WUI is a client of RODA Core Services.

RODA Core Services are responsible for carrying out more complex tasks than the ones offered by Fedora Commons such as the complete set of procedures that compose the ingest workflow, querying the repository in more advanced and abstract ways and carrying out administrative functions on the repository.

For performing format migrations the Core Services rely on RODA Migration services. RODA Migration services are several web services that perform format conversions on dedicated remote servers.

4. Future and current work

The RODA team is always engaged in addressing new challenges, and to go beyond the state-of-the-art. RODA is now part of the SCAPE project, an international project with several European national libraries and universities with the objective of enhancing the state of the art of digital preservation in three ways: by developing infrastructure and tools for scalable preservation actions; by providing a framework for automated, quality-assured preservation workflows; and by integrating these components with a policy-based preservation planning and watch system. Currently, the team is focusing in scalability, automatic monitoring of the world and preservation planning integration.The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

5. REFERENCES

[1] Consultative Committee for Space Data Systems (CCSDS), *Reference Model for an Open Archival Information System (OAIS)*. Consultative Committee for Space Data Systems (CCSDS), 2002.

[2] "Metadata Encoding and Transmission Standard (METS) Official Web Site." [Online]. Available: http://www.loc.gov/standards/mets/. [Accessed: 09-Aug-2011].

[3] "EAD: Encoded Archival Description Version 2002 Official Site (EAD Official Site, Library of Congress)." [Online]. Available: http://www.loc.gov/ead/. [Accessed: 09-Aug-2011].

[4] PREMIS Editorial Committee, *Data Dictionary for Preservation Metadata: PREMIS version 2.0*. 2008.

A non-proprietary RAID replacement for long term preservation systems

Samuel Goebert
Hochschule Darmstadt
University of Applied Science
Darmstadt, Germany
samuel.goebert@bigcurl.de

Alain Sarti
Hessen Main State Archive
Wiesbaden, Germany
alain.sarti@hhstaw.hessen.de

1. INTRODUCTION

Disk-based storage, as suggested by Rosenthal et al., is the de facto standard for long term storage solutions [1]. Established by Patterson et al., RAID-based systems are a basic building block and have been a common best practice for building large scale storage systems [2].

Increasing disk failure rates, as experienced by Pinheiro et al. and proprietary ways to access the file system render the benefits of a RAID-based system questionable for longterm preservation [3].

A system with a RAID-level of 6 loses all its data if three disks are unavailable at the same time. Otherwise the lost content can be recovered by replacing one or both of the unavailable disks (with either a new one or a hot spare one).

We developed a RAID-free storage system that is able to replace RAID as a fundamental building block. The approach named NRN (No-RAID-Necessary) distributes a configurable number instances of data over a number of drives. To experience data loss all disks which hold an instance of a file have to fail. Even in this case only the data on those disks is lost. The data on the other discs is still accessible.

2. OVERVIEW

In large scale network storage systems like Amazon S3 and LOCKSS, the concept of treating a file as a whole object is part of the strategy against data loss. Splitting the file into chunks and putting them on different nodes in the network raises the overall speed, while accessing the file but also raises the number of machines necessary to fully recover a file [6] [7].

While it is possible to recalculate missing chunks of a file if redundant information is added, it still depends on the algorithm, how many of the chunks have to be intact to recalculate the missing chunks in the file. In a one chunk per node distribution strategy, the number of nodes that have to be intact to fully retrieve a file is determined by the algorithms ability to recover the file.

This is in contrast to a whole object approach, where one node holds a complete copy of a file. While this approach takes up more storage space since multiple copies of a file are stored in the system, only one node is needed to fully retrieve a copy. This makes the overall system robust against node failures.

While this approach is widely used for nodes in a network, the nodes themselves follow a different pattern for storing the files on disk. In large systems up to 48 hard drives are used per machine to form a node in the system. In most cases a RAID system is used to let the drives appear as a single large volume to the software that stores files into these disks.

Instead of taking the same approach as the network level and translate it to the disks instead of nodes, the files are split up and distributed over several disks with all the disadvantages that are avoided on the network level. A single failing drive can take down a complete node, which might result in many network traffic since the minimal number of copies of a file is enforced by the managing system.

What NRN does is taking the lessons learned from the network level and applies them to the node level. Individual disk fulfill the same role as nodes on the network and store full copies of the file on more than one disk. If a drive fails, only the missing content from the drive has to be replicated. Replication on the local bus happens with maximum bandwidth provided by the drives and does not utilize CPU cycles while copying data.

Also only one disk is affected during the recovery stage and not the whole system, as it would be the case in a RAID system where the missing data is recalculated from the remaining disks. Every workflow that might be enforced during the boot process like a fsck disk check up can be started in parallel on all disks which greatly improves boot up time of a node.

3. APPROACH

In the presented approach, several consumer grade disks are connected individually to a system, which, in contrast to a RAID system, are not logically combined to form one big drive. The disks are accessed through a thin software layer,

iPRES2011, Nov. 1–4, 2011, Singapore.

which is responsible for replicating incoming files over several disks. Several instances of a file establish the basis to achieve high availability for the files and durability against disk failures.

The abstraction through which the overlaying software access NRN is provided by FUSE *http://fuse.sourceforge.net*. FUSE allows a filesystem implementation in user space without coding directly in the kernel. The resulting filesystem hides the complexity of dealing with all disks individually and provides a single folder interface for the software that wants to store data. This makes it also possible to utilize the storage method also with software that is not optimized for the usage of more than one disk.

NRN always returns one of the available data instances when a client requests a file. This ensures that a file can be retrieved, even if the system has detected a disk failure and while redistributing the lost data. In case of a disk failure, it tries to comply with the predefined number of instances by recreating them on a hard disk which does not contain one already.

Recovery time in a NRN system depends on the amount of data on the failed disk, not on total system capacity or even individual disk capacity. If remaining total system capacity allows it, the system does not need a hot spare or replacement disk to start issuing new instances of the lost data onto the remaining drives.

The problem of distributing the instances onto the disks is solved by using one dimensional bin packing problem algorithms. For this purpose Lee et al. provided a first fit algorithm [4]. We identified two approaches, which suits long term archives best. They only vary by the number of disks available to the algorithm.

In the first approach we put a strong emphasis on high availability. Data is distributed equally onto all available disks. The drive with the lowest total capacity stores a new file. Due to the fact that only a fraction of files have to be restored, the recovery time from a drive failure is minimal. A higher number of disks means better protection against total data loss.

The second approach puts the emphasis on growing the capacity as needed. Disks are filled one by one. Initially only the minimum number of disks have to be attached to the system. If full capacity is reached, more disks are attached to the system to expand the overall capacity. This approach enables to start with a small upfront investment and only add drives when they are really needed.

4. CONCLUSION

Our research revealed that by replacing the RAID components with a system running NRN we have to accept a lower space usage efficiency and throughput. It is possible to keep most benefits from the RAID approach like robustness against individual disk failure and hot spare disks to lower maintenance reaction times. In addition we removed the proprietary file system, decoupled the system recovery time from the total disk capacity and lowered the probability of total data loss.

5. REFERENCES

[1] D. S. H. Rosenthal, M. Roussopoulos, TJ Giuli, P. Maniatis, and M. Baker. Using hard disks for digital preservation. In *Imaging Sci. and Tech. Archiving Conference*, 2004.

[2] D. A. Patterson, G. Gibson, and R. H. Katz. A case for redundant arrays of inexpensive disks (raid). In *SIGMOD88 International Conference On Management of Data*, SIGMOD '88, pages 443, Chicago, IL, USA - June 01 - 03, 1988. ACM.

[3] E. Pinheiro, W.-D. Weber, and L. A. Barroso. Failure trends in a large disk drive population. In *Proceedings of the 5th USENIX conference on File and Storage Technologies*, pages 2–2, Berkeley, CA, USA, 2007. USENIX Association.

[4] C. C. Lee and D. T. Lee. A simple on-line bin-packing algorithm. *JACM*, 32:562–572, July 1985.

[5] P. Constantopoulos, M. Doerr, and M. Petraki. Reliability modeling for long term digital preservation abstract. In *9th DELOS Network of Excellence thematic workshop "Digital Repositories: Interoperability and Common Services"*, Foundation for Research and Technology, Hellas (FORTH), Heraklion, Crete 11-13 May, 2005

[6] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, and W. Vogels. Dynamo: amazon's highly available key-value store. *SIGOPS Oper. Syst. Rev.*, 41:205–220, October 2007.

[7] V. A. Reich and D. S. H. Rosenthal. Lockss: Building permanent access for e-journals â practical steps towards and affordable, cooperative, e-preservation, and e-archivingprogram. In *ELPUB*, 2003.

An Open-Source System for Automatic Policy-Based Collaborative Archival Replication

Thu-Mai Christian
& Jonathan Crabtree
University of North Carolina
Odum Institute
jonathan_crabtree@unc.edu

Nancy McGovern
University of Michigan
ICPSR
nancymcg@umich.edu

Micah Altman
Harvard University
IQSS
micah_altman@harvard.edu

ABSTRACT
In this poster, we provide an overview of the SafeArchive system and describe how a curator can use the tools to generate an archival policy schema and monitor compliance. Also, the poster details the technical implementation of the SafeArchive system including the policy schema, how information used in the auditing process is obtained from a set of LOCKSS peers without modifying the LOCKSS trust model or configuration, and the organization of SafeArchive software components.

Categories and Subject Descriptors
H.3.4 [**Systems and Software**]: Distributed Systems Audit

General Terms
Management, Measurement, Documentation, Performance, Reliability, Legal Aspects, Verification

Keywords
Audit, Open-Source, Policy, LOCKSS, TRAC, Preservation, Archive

1. INTRODUCTION
Verified geographically-distributed replication of content is an essential component of any comprehensive digital preservation plan. This requirement has emerged as a necessity for recognition and certification as a trusted repository. As embodied in Trustworthy Repositories Audit & Certification (TRAC) [1] and the subsequent TRAC-based ISO 16363 Audit and Certification of Trustworthy Digital Repositories, and in other best practices, an organization must have a managed process for creating, maintaining, and verifying multiple geographically distributed copies of its collections in order to be fully trusted.

The LOCKSS (Lots of Copies Keep Stuff Safe) [2] system has been widely adopted by libraries and archives for replication and preservation. As a collaborative effort of Data-PASS partners (ICPSR, Roper Center, University of Connecticut, Odum Institute and IQSS), the SafeArchive system has been developed to extend LOCKSS capabilities by making distributed replication easier for curators and automating compliance with formal replication and storage policies. This innovation provides the auditability and reliability of a top-down replication system with the resilience of a peer-to-peer model.

iPRES2011, Nov. 1–4, 2011, Singapore.

2. THE SYSTEM
2.1 Overview of the SafeArchive System
SafeArchive is described in more detail in [3] and is based on a prototype [4] developed by the Data-PASS partners [5, 6], and funded by the Library of Congress. This prototype established feasibility and the core operational use cases for the system. The SafeArchive system has been completely rewritten and redesigned for production use.

Abstractly, the system is designed to create a virtual overlay network on top of a peer-to-peer replication network that supports provisioning, monitoring, and TRAC/ISO 16363-based auditing.

Operationally, users of the system can perform the following functions, as illustrated in figure 1:

- Analyze any LOCKSS network;
- Check that collections are replicated, valid, and up-to-date;
- Create formal replication policies;
- Replicate content from web sites or digital repository systems;
- Audit the network for current and historical TRAC/ISO 16363 compliance; and
- Automatically manage and repair a LOCKSS network based on a specified replication policy.

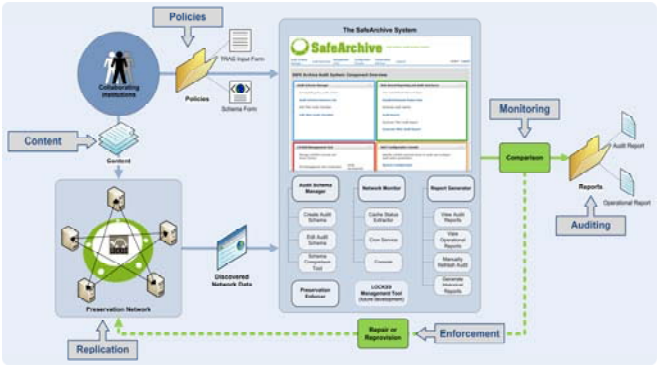


Figure 1. Abstract diagram of system functions and roles

The SafeArchive system is designed to collaborate with the Dataverse Network® [7] system. Curators who store content in a Dataverse can easily expose content for replication by LOCKSS and SafeArchive through a simple graphical interface.

Institutionally, the SafeArchive enables memory institutions and preservation collaborations to formalize their replication policies and inter-archival replication commitments; represent these replication policies in machine-readable form; and to continuously audit any set of public or private LOCKSS hosts for policy compliance. The SafeArchive system is open source and available at: http://www.safearchive.org.

2.2 Using the SafeArchive System

Generally speaking, the system coordinates six activities:

1. Collaborating institutions agree on a replication policy. This records the resource commitments, descriptions of the collections to be preserved, and desired replication guarantees.
2. Institutions make collections of content ("archival units") available through the web (e.g., as web pages or through the Dataverse Network®).
3. LOCKSS caches harvest the collections from their original source repositories using standard protocols such as HTTP or OAI-PMH.
4. SafeArchive monitors the network, assesses it against the stated replication policy, and produces an audit trail. The system also alerts collaborators when formal policies are not being met.
5. SafeArchive produces an audit trail of operational and audit reports.
6. The SafeArchive will also coordinate harvesting of the LOCKSS caches by "inviting" members of the network to harvest content that is under-replicated. This will be used to automatically configure a network based on a policy schema to reconfigure and repair the network as the number of participating caches, collections and institutions changes intentionally or unintentionally.

The SafeArchive system is designed to give curators the ability to easily define preservation policy, examine the content of the preservation network, and generate regular audit reports that support TRAC/ISO 16363 compliance. All changes to the policy schema instance and the machine-readable audit reports are versioned and stored permanently—so that a complete history of compliance is preserved.

3. SUMMARY

The SafeArchive system provides a way to ensure that replicated collections are both institutionally and geographically distributed while allowing for the development of increasingly measurable and auditable trusted repository requirements. Designed as a virtual overlay network on LOCKSS, the system provides the auditability and reliability of a top-down replication system with the resilience of a peer-to-peer model. This enables any library, museum, or archive to audit the replication of their collections across an existing LOCKSS network in compliance with documented archival policies. It also allows groups of collaborating institutions to automatically and verifiably replicate each others' content consistent with a set of expressed commitments stored in machine readable XML based policies. The result is that archives can more easily collaborate to preserve content through geographically and institutionally distributed replication, which mitigates technical and organizational threats to preservation.

The project is in its second year of development and the first official version 1.0 of the system has been released. The system is being field-tested, and optimizations from those experiences are being incorporated into version 2.0 that is slated for release in early 2012.

4. ACKNOWLEDGEMENTS

The project is a collaborative effort of the Data-PASS Partners: The International Consortium for Political and Social Research, University of Michigan; The Roper Center for Public Opinion Research, University of Connecticut; the Howard W. Odum Institute at the University of North Carolina at Chapel Hill; the National Archives and Records Administration; and the Institute of Quantitative Social Science, Harvard University. It is managed through the Institute of Quantitative Social Science, and works in collaboration with the LOCKSS project at Stanford University.

The project is sponsored by the Institute of Museum and Library Services (IMLS), under award #LG-05-09-0041-09.

5. REFERENCES

[1] RLG-NARA Task Force on Digital Repository Certification. (2007). Trustworthy Repositories Audit and Certification (TRAC): Criteria and checklist (version 1.0). Chicago, IL: Center for Research Libraries. Retrieved from http://www.crl.edu/PDF/trac.pdf

[2] Reich, V. & Rosenthal, D. (2001). LOCKSS: A permanent web publishing and access system. *D-Lib Magazine*, *7*(6). Retrieved from http://www.dlib.org/dlib/june01/reich/06reich.html

[3] Altman, M., & Crabtree, J. (2011). Using the SafeArchive System : TRAC-based auditing of LOCKSS. *Archiving 2011* (pp. 165-170). Society for Imaging Science and Technology. doi:ISSN 978-0-89208-294-0

[4] Atman, M., Beecher, B., Crabtree, J., Andreev, L., Bachmann, B., Buchbinder, A., Burling, S., King, P., & Maynard, M. (2009). A prototype platform for policy-based archival replication. *Against The Grain*, *21*(2), 44-47.

[5] Altman, M., Adams, M., Crabtree, J., Donakowski, D., Maynard, M., Pienta, A., & Young, C. (2009). Digital preservation through archival collaboration: The Data Preservation Alliance for the Social Sciences (Data-PASS). *The American Archivist*, *72*(1), 169-182.

[6] Gutmann, M., Abrahamson, M, Adams, M.O., Altman, M., Arms, C., Bollen, K., Carlson, M., Crabtree, J., Donakowski, D., King, G., Lyle, J., Maynard, M., Pienta, A., Rockwell, R, Timms-Ferrara L., & Young, C. (2009). From preserving the past to preserving the future: The Data-PASS project and the challenges of preserving digital social science data. *Library Trends*, *57*(3), 315-337.

[7] Crosas, M. (2011). The Dataverse Network®: An open-source application for sharing, discovering and preserving data. *D-Lib Magazine 17*(1/2). Retrieved from http://www.dlib.org/dlib/january11/crosas/01crosas.html

TOTEM: Trusted Online Technical Environment Metadata - A Long-Term Solution for a Relational Database / RDF Ontologies

Dr Janet Delve
Future Proof Computing Group
School of Creative Technologies
University of Portsmouth
004423 9284 5524
Janet.Delve@port.ac.uk

Dr Leo Konstantelos
Future Proof Computing Group
School of Creative Technologies
University of Portsmouth
004423 9284 5491
Leo.Konstantelos@port.ac.uk

Dr Antonio Ciuffreda
Dr David Anderson
Future Proof Computing Group
School of Creative Technologies
University of Portsmouth
004423 9284 5491
Antonio.Ciuffreda@port.ac.uk

ABSTRACT

For emulation and other preservation actions, metadata is needed to describe the technical environment (operating system, related software libraries, hardware etc.) in which a given file or item of software can be rendered. This paper delineates an enhanced entity attribute relationship model suitable as a basis for a database (relational, object-relational or object-oriented), or for a RDF ontology. This core data model covers the X86, Apple II and Commodore 64 (C64) hardware architectures, as well as games consoles. The model is currently instantiated as a MySQL database with accompanying API, plus a PHP-based browsing system. Data population, and user evaluation are also discussed. The model is extensible over the long term and is compatible with OAIS and PREMIS version 2.

1. INTRODUCTION

A state-of-the art survey [2] for the KEEP project examined in depth the existing technical environment metadata, and found there to be some preparatory work on which to build, but no extant, completed data models / schemas available explicitly for this purpose. PREMIS 2 confirmed this finding, and provided guidelines for technical environment metadata in either database or ontology format [5]. An Enhanced Entity Attribute Relationship conceptual model was thus chosen for KEEP as it provided a basis for either format.

2. DATA MODELS

The core version of the TOTEM Enhanced Entity Relationship Diagram (EERD) is generic, and was created via a bottom-up approach using catalogue data for a PDF file, a multimedia encyclopedia and a console game. The catalogue data held some technical environment information including an initial range of hardware such as 'a multimedia PC' and 'an Apple II'. Three of the publications were on media carriers: CD-ROMs, a 5 ½" floppy disk and a games cartridge. See [3] for full details and EERDs.

A particular feature of this data model is its granularity. For a PDF file, it is vital to know its version, as specific software is required to render a specific PDF version. So a PDF version 1.4 file would run on Adobe Acrobat version 5.x software on a Mac OS version X 10.5.6 (9G55) Operating System (OS) on a MacbookPro 5.1 hardware platform. In contrast, PDF/X files incorporate a Graphics Art Technologies standard and need a software package such as CorelDRAW 11 to render them. Another vital feature of many technical environments is software libraries (for example .DLL files) that the OS might need. Note that further extensions of the EERD will embrace: software patches; system libraries, plug-ins, fonts; plugged hardware devices with their corresponding drivers (together with driver versions) and BIOS revision. This generic model covers a variety of PC architecture together with the software and operating systems that run on them. Computer games running on a PC such as the C64 can be modeled here, but the model for the Games Console in is different in structure to a typical computer, given that the computer chip resides in the console OS. Much thought was given to modeling the plethora of appendages used with games consoles (i.e. joysticks). Concerning the games controllers, generic attributes were defined for a particular version of a game running on a particular games console as being either analogue or digital, and these are then further refined in terms of planes and degrees of movement etc.

Having created these conceptual data models, research was carried out to establish the best way of implementing them. A method embracing linked data was identified in the initial survey [2] and Dublin Core[1] as a good way of achieving semantic interoperability which is achieved through the use of a common RDF format which facilitates the gathering of information via linked data clouds [1]. RDF ontologies would thus appear to be an ideal vehicle for the KEEP technical environment metadata. However, the Planets project[2] had carried out extensive XCL characterization work that included the development of a raft of ontologies [4], but it was pointed out that OWL Protégé had some shortcomings for the general user, and Excel spreadsheets were used instead to house these ontologies:

"A solution may be for non-OWL experts to develop class structures by hierarchically organising relevant concepts in a spreadsheet, and having an OWL expert or software developer

iPRES2011, Nov. 1–4, 2011, Singapore.

[1] http://dublincore.org/metadata-basics/

[2] http://www.planets-project.eu/

develop a script for transforming this spreadsheet into the RDF/OWL language. Such a procedure has been followed manually (not involving scripts) in developing the initial PC ontology, which proved to be a lot faster than building it as an OWL ontology in Protégé, because of the large numbers of classes and individuals involved. This may be an efficient way of developing ontologies in future within the digital preservation community." Collaboration with the University of Cologne (Universität zu Köln) is underway to convert the EERDs into RDFs for the software and hardware classes [8].

3. THE TOTEM DATABASE

The conceptual models outlined above have formed the basis for the specification of the TOTEM database Twenty five entities have been modeled, comprising more than 130 elements and their relationships, in a fully normalized structure. Three distinct technical environments are currently supported in the logical data model: the PC architecture, the Commodore 64 architecture and console gaming platforms. It is possible however to represent additional environments (e.g. Apple II or Acorn) in the future, by following the specifications in the conceptual models. The physical model was developed as a MySQL database, accessible as part of KEEP emulation services as well as to general DP users.

Three distinct user roles have been identified: end users, metadata data administrators and metadata database administrators. Database administration is managed via the phpMyAdmin[3] open source tool, currently deployed over an Apache web server. Interaction between the database and end users / data administrators is made via a database application that acts as a front-end and browsing system.

The browsing system, implemented in PHP, allows access to browsing and searching the TOTEM database through a simple interface currently providing three types of search functionality: simple search; advanced search; and compatibility search. In the compatibility search, the user can explore: Software types compatible with a specified file type and version; Software libraries compatible with a specified software type and version; Operating systems compatible with a specified software type and version; and Hardware types compatible with a specified operating system and version.

The greatest challenge **(is)** the process of identifying, populating and maintaining the resource with accurate, pertinent and up-to-date data. TOTEM currently holds PC-related technical metadata, Commodore 64- and Console Game-related metadata. The sheer volume of data – alongside the process of corroborating their accuracy and hence usefulness – clearly indicates that this task cannot be single-handedly undertaken by a sole institution or a sole project. Long-term sustainability of the TOTEM resource and continuing adoption of the model and deriving schema necessitate equally continuing support from user communities. Having identified these caveats, a number of potential solutions are being explored, including collection of data from product documentation and developer blogs, and Crowdsourcing data population by making the database accessible to relevant communities.

The resource will be integrated with the suite of tools provided under the aegis of the Open Planets Foundation (OPF)[4] registry ecosystem. [6] sets out the vision for a new registry for digital preservation, or a "registry ecosystem", which will build on linked data in order to create an interconnected collection of existing (and future) information registries that currently exist in isolation. Although this can be a sustainable solution, the risk of erroneous/contradictory information being inserted, with varying degrees of detail and granularity **still exists**. The OPF registry ecosystem envisages countering this problem by promoting the Crowdsourcing path and by introducing tools that allow institutions to set their own confidence levels on representation information in registries [7]. To conclude, comprehensive user evaluation for TOTEM is almost complete and feedback received so far indicates that this is a useful weapon in the DP armory. Detailed plans for future improvements are already mapped out to make this a robust, versatile, scalable and shareable tool.

[3] http://www.phpmyadmin.net/

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

4. ACKNOWLEDGMENTS

The Keeping Emulation Environments Portable (KEEP) Project is co-financed by the European Union's Seventh Framework Programme for research and technological development (FP7), Grant Agreement number ICT-231954.

5. REFERENCES

[1] Anderson, D., Delve, J., and Pinchbeck, D. 2010. Toward A Workable Emulation-Based Preservation Strategy: Rationale and Technical Metadata. *The New Review of Information Networking* 15, 2 (Nov. 2010), 110-131. DOI=http://dx.doi.org/10.1080/13614576.2010.530132.

[2] Anderson, D., Delve, J., Pinchbeck, D., and Alemu, G. A. 2009. *Preliminary document analyzing and summarizing metadata standards and issues across Europe*. KEEP Technical Report D3.1. URL= http://www.keep-project.eu/ezpub2/index.php?/eng/content/download/4124/20617/file/KEEP_WP3_D3.1.pdf

[3] Delve, J., Ciuffreda, A., and Anderson, D. 2010. *Documents describing meta-data for the specified range of digital objects, as well as requirements and design for the browsing system and user interface of the Emulation Framework.* KEEP Technical Report D3.2.

[4] Montague, L., Nicchiarelli, E., Matthezing, H., Kummer, R., Puhl, J., & Roberts, B. (2010a). *Planets components for the extraction and evaluation of digital object properties*. Planets Technical Report D23B. URL=http:// www.planets-project.eu/docs/reports/Planets_PC3-D23B(DOPWGreport).pdf

[5] PREMIS Working Group, OCLC, & RLG. 2008. *PREMIS data dictionary for preservation metadata Version 2.0.* Washington, DC: Library of Congress. URL=http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

[6] Roberts, B. 2011. *A New Registry for Digital Preservation: Conceptual Overview*. Open Planets Foundation Technical Report version 1.1. URL= http://www.openplanetsfoundation.org/new-registry-digital-preservation-conceptual-overview

[7] Tarrant, D., Hitchcock, S. and Carr, L. 2009. Where the Semantic Web and Web 2.0 meet format risk management: P2 registry. In *Proceedings of the 6th International Conference on Preservation of Digital Objects* (San Fransisco, CA, October 5-6, 2009). CDL, San Fransisco, CA, 187-193. DOI= http://escholarship.org/uc/item/8525r8cn Trust levels are discussed on p17

[8] Thaller, M. 2009. The eXtensible Characterisation Languages - XCL. Verlag Dr. Kovac, Hamburg.

Corporate Recordkeeping: New Challenges for Digital Preservation

Gillian Oliver
School of Information Management
Victoria University of Wellington
Wellington, New Zealand
Phone (+ 64) (0)4 463 7437
Gillian.Oliver@vuw.ac.nz

Fiorella Foscarini
Faculty of Information,
University of Toronto
Toronto, Canada
Phone (+1) 416-978-8295
fiorella.foscarini@utoronto.ca

ABSTRACT
In this paper, we describe an innovative approach to the challenges associated with managing corporate records in the digital environment. Issues and problems with the use of EDRMS are well documented but alternatives are not yet mature enough for workplace implementation. The recordkeeping functionality of Microsoft SharePoint is disputed by practitioners, but this enterprise content management system appears to be emerging as a default solution to manage records. Applying genre theory in the configuration of SharePoint will assist records managers in negotiating shared understanding with their information technology colleagues which is essential in order to achieve digital preservation objectives.

Categories and Subject Descriptors
D.2.9 [**Software Engineering**]: Management – *Software configuration management*.

General Terms
Management, Design, Theory.

Keywords
MS SharePoint, EDRMS, content type, genre theory.

1. INTRODUCTION
The challenges of ensuring that born digital information is not only accessible for as long as required, but also can be trusted for evidential purposes have long been the focus of study by the records community. Up until recently the predominant solution proposed has been electronic document and records management systems (EDRMS). However, EDRMS implementations have not been without problems and so alternative approaches are being investigated.

In the meantime, Microsoft SharePoint is rapidly achieving market dominance, and in many cases may be the only option available for corporate recordkeeping. A central feature of MS SharePoint is Content Type, which can be considered from the perspective of genre theory. This paper suggests that genre theory provides an innovative approach to capturing the context of records creation and use, and as such, it may be usefully drawn on for the analysis and identification of Content Type.

iPRES2011, Nov. 1–4, 2011, Singapore.

2. DIGITAL RECORDKEEPING SYSTEMS
Research conducted by the international archival and records management community in the last two decades unanimously recommends that recordkeeping functionalities be implemented within offices with active records [3]. This recommendation has been widely adopted by organizations of all types across the globe. Electronic recordkeeping systems have been developed by a variety of different vendors and may be referred to by a range of acronyms (such as, EDMS, ERMS, ECMS, EDRMS). In some jurisdictions, these systems are required by law, but the challenges faced in implementation [9, 17] have motivated research to identify best practice and success factors [7, 16]. A key concern that has emerged from the literature is reluctance on the part of users to engage with EDRMS [13].

Enterprise content management in the form of SharePoint is being presented by Microsoft as a solution that can encompass all digital content, including records [4]. However, the extent to which MS SharePoint can be considered a recordkeeping system is hotly debated by the recordkeeping professional community [8]. A specific issue relates to the critical role of *Content Types* in MS SharePoint. Content Types are a new concept for the records community, and furthermore they have been singled out as one of the major disadvantages associated with using the MS SharePoint records centre. This is because they are perceived as being difficult to administer and understand. As Lappin puts it, "content types … are more powerful than folders, but they are also more complex for you as an administrator to set up and maintain, and for your colleagues to understand and use" [8]. In the light of this strongly worded warning, it seems very important indeed for records managers to get to grips with the concept of Content Type. The official Microsoft definition of Content Types is as follows: "a reusable collection of metadata (columns), workflow, behavior, and other settings for a category of items or documents in a Microsoft SharePoint Foundation 2010 list or document library" [10]. Or, features that enable identifying what a document *is* through what it *does*. This implies a much more multi-faceted approach to Content Type than its name suggests. This is significant because it shows potential for interpretation of Content Type as a genre-like concept.

2. GENRE & CONTENT TYPE

Genre can be defined as a socially recognised communication norm which can range from speech acts to text messages, and has been the subject of research in a number of different disciplines including communication and information studies [1, 15]. Of particular interest and relevance to the records management community is research into genre in organisational contexts [19] and the concept of genre system, which ensures a holistic view of communicative actions [2, 14, 15]. Introducing a genre approach to digital recordkeeping represents exciting potential for a new way of thinking for records managers. The emphasis in current practice is on determining the functions that are carried out by organisations, and from there identifying the records that are being created as a consequence of transactions carried out to support those functions. This is clearly stated in the international standard for records management ISO15489 [6]; but actually identifying and defining functions is fraught with difficulty and ambiguity [5]. A genre perspective on the information created and maintained within organisations allows for a *situated approach* where the functional and social contexts that enabled the emergence of specific patterns of communicative actions become apparent [12, 18]. Some of the insights offered by genre theory may help furnish a common ground to foster the development of shared understanding between records and information technology professionals. This common ground is essential if digital preservation goals are to be achieved.

Each Content Type involves a number of attributes, some rather generic (e.g., properties, metadata, custom features), others more specific (e.g., workflows, information management policies, document templates) and may refer to any kinds of information objects (e.g., list items, documents, folders, photos, videos, blogs) [10, 11]. Because they seem to have unlimited coverage, Content Types tend to confuse SharePoint users. We suggest that the use of Content Type would be facilitated by relating it to the concept of genre as "typified communicative action" [18]. In particular, Yates and Orlikowski's dimensions of communicative action (i.e., What, How, Who, When, Why, Where) [14] would help classify the attributes involved in a more consistent way. Content Type mixes up elements of form (templates), substance (the name itself 'content' type), and context (workflow), while genre theory clearly says that the *action* accomplished by the genre in a specific situation is the criterion to categorize classes of information objects. Action is also inextricably linked to records, which are defined in ISO15489 as "information created, received and maintained as evidence and information by an organization or person, in pursuance of legal obligations or in the transaction of business" [6]. In other words, records have to be associated with 'doing something'. Genre theory appears to provide a disciplinary appropriate lens for records managers to view and define Content Type in such a way as to ensure that context, as well as content, is taken into account, thus maximizing recordkeeping functionality.

3. NEXT STEPS

Further work is needed to test the assumptions made on this paper. One approach will be to survey organizations with similar functions currently using SharePoint to collect data to show how Content Type is currently being interpreted and used. The next stage will be to define a set of Content Types appropriate to each sector, using Yates and Orlikowski's dimensions of communicative action. The resulting set can then be tested as a prototype in work environments. At this early theoretical stage, however, we can conclude that genre theory offers exciting possibilities for new approaches to the challenges of ensuring that digital records can be maintained for as long as they are required and a way out of the current EDRMS dilemma.

4. REFERENCES

[1] Andersen, J. 2008. The concept of genre in information studies. *Annu. Rev. Inform. Sci.* 33, 9-67.

[2] Bazerman, C. 1994. Systems of genres and the enactment of social intentions. In *Genre and the new rhetoric*, A. Freedman, P. Medway, Eds. Taylor and Francis, London, 79-101.

[3] Bearman, D. 2006. Moments of risk: Identifying threats to electronic records. *Archivaria* 62, 15-46.

[4] Duguid, R. (2010). Introducing records management in SharePoint 2010. DOI= http://blogs.msdn.com/b/ecm/archive/2010/02/13/introducing-records-management-in-SharePoint-2010.aspx.

[5] Foscarini, F. 2010. Understanding the context of records creation and use: 'Hard' versus 'soft' approaches to records management. *Arch. Sci.* 10 (4), 389-407.

[6] International Organization for Standardization. 2001. ISO 15489 Information and documentation – Records management. ISO, Geneva

[7] JISC InfoNet. 2009. Implementing an electronic document and records management (EDRM) system. DOI= http://www.jiscinfonet.ac.uk/InfoKits/edrm

[8] Lappin, J. 2010. Is there a sustainable and scalable records management model in SharePoint 2010? DOI= http://www.aiim.org/community/blogs/expert/Is-there-a-sustainable-and-scaleable-records-management-model-in-SharePoint-2010

[9] Maguire, R. 2005. Lessons learned from implementing an electronic records management system. *Rec. Manage. J.* 15 (3), 150-157.

[10] Microsoft. 2010. Introduction to content types. DOI= http://msdn.microsoft.com/en-us/library/ms472236.aspx.

[11] Microsoft. 2010. TechNet Library. Content type and workflow planning (SharePoint Server 2010). DOI= http://technet.microsoft.com/en-us/library/cc262735.aspx.

[12] Miller, C.R. 1984. Genre as social action. *Q. J. Speech* 70 (2), 151–167.

[13] Oliver, G. 2011. *Organisational culture for information managers*. Chandos Press, Oxford.

[14] Orlikowski, W. and Yates, J. 1994. Genre repertoire: the structuring of communicative practices in organizations. *Adm. Sci. Quart.* 39 (4), 541-574.

[15] Osterlund, C. 2007. Genre combinations: a window into dynamic communication practices. *J. Manage Inform. Syst.* 23 (4), 81-108.

[16] Reed, B. 2008. Service oriented architectures and recordkeeping. *Rec. Manage. J.* 18 (1), 7-20.

[17] Smyth, Z.A. 2005. Implementing EDRM: has it provided the benefits expected? *Rec. Manage. J.* 15 (3), 141-149.

[18] Yates, J. and Orlikowski, W. 1992. Genres of organizational communication: a structurational approach to studying communication and media. *Acad Manage Rev.* 17 (2), 299-332

Preserving Change: Observations on Weblog Preservation

Yunhyong Kim
Humanities Advanced Technology and Information Institute
University of Glasgow
Glasgow, UK
Yunhyong.Kim@glasgow.ac.uk

Seamus Ross
Humanities Advanced Technology and Information Institute
University of Glasgow, Glasgow, UK
&
Faculty of Information
University of Toronto
Toronto, Ontario, Canada
seamus.ross@utoronto.ca

ABSTRACT
In this article, we revisit concepts introduced within the digital preservation literature, such as Open Archival Information System (OAIS) reference model, and Preservation Metadata Implementation Strategy (PREMIS), to examine their continued applicability to the preservation of dynamic web content such as weblogs.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – Standards.

General Terms
Management, Design, Human Factors, Standardization, Theory.

Keywords
digital preservation, digital curation, designated community, authenticity, intellectual entity, archive, web archive, blog, weblog

INTRODUCTION
Current preservation approaches tend to be largely data object oriented, relying on the notion that data can be reasonably reduced to a manageable discrete set of objects accompanied by formal syntactic, semantic and pragmatic attributes that constitute the original object's content and characteristics necessary for validating authenticity, managing rights, and enabling access and use (e.g. see [1], [6]). Now, the dynamic web environment (e.g. blogs, wiki, networking platforms) enables us to capture data objects at finer levels of communicative granularity. Continuing to capture each of these bits as a discrete entity/object imposes independent object identities on pieces of information that, in the past, would have only been considered to have meaning as part of the whole intellectual process. It may be time to re-examine the established approaches to determine whether they are still valid in the context of web archiving initiatives (e.g. the Minerva Project[1], Internet Archive[2], UK Web Archive[3], Arcomem[4], BlogForever[5], Memento Project[6], LiWA[7]) that have been increasingly trying to create solutions for social media archival situations.

iPRES2011, Nov. 1–4, 2011, Singapore.

OAIS: A BRIEF SUMMARY
The Reference Model for an Open Archival Information System (OAIS) [1] is a conceptual model for a preservation-aware archival system developed by the Consultative Committee for Space Data Systems (CCSDS) (accepted as an ISO standard in 2003[8]). It has been adopted by several well-known preservation projects in recent years (e.g. CASPAR[9], SHAMAN[10], SHERPA DP2[11] and the Planets Interoperability Framework [9]). To be compliant to the model (see [1]), "the OAIS must: 1) negotiate for and accept appropriate information from information producers; 2) obtain sufficient control of the information needed to ensure long-term preservation; 3) determine which communities should become the Designated Community and, therefore, should be able to understand the information provided; 4) ensure that the information to be preserved is independently understandable to the Designated Community; 5) follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original, or as traceable to the original; and, 6) make the preserved information available to the Designated Community."[12].

PREMIS DATA MODEL
The PREMIS (Preservation Metadata: Implementation Strategies) working group was sponsored by OCLC Online Computer Library

[1] http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html

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

[3] http://www.webarchive.org.uk

[4] http://www.arcomem.eu

[5] http://www.blogforever.eu

[6] http://www.mementoweb.org/

[7] http://www.liwa-project.eu/

[8] ISO/DIS 14721

[9] http://www.casparpreserves.eu

[10] http://shaman-ip.eu/shaman/

[11] http://www.sherpadp.org.uk/sherpadp2.html

[12] This content from [1] has been condensed to save space.

Center and Research Libraries Group (RLG), to develop a core set of preservation metadata applicable to a wide range of digital preservation contexts. The resulting standard [6] was intended to comply with the OAIS model (Section 2.1), while targeting metadata that capture preservation processes, such as the preservation level associated with an object. While descriptive and technical metadata are also key concepts in the standard, PREMIS recommends the use of previous standards to meet requirements for these, focusing on preservation levels and processes, rights, and object properties and relations to be preserved. A large amount of the effort in PREMIS remains with object modeling. While notions of agents, events and rights are discussed within the standard, detailed information is not provided. The model relies on the concept of an **intellectual entity** as a single intellectual unit to be managed within the archive.

OBSERVATIONS ON WEB ARCHIVING

While many web archives have claimed compliance with the OAIS model (Section 2.1), this can be accepted only on the most generous terms: 1) while access can be blocked, there is almost never any explicit negotiation between information producers and existing web archives: the information is obtained through procedures for "copying the website" [7]; 2) the lack of negotiation means that the archive's rights to manipulate harvested pages for preservation purposes becomes ambiguous, and introduces an unpredictable gap between the archive's "authentic copy" and the material at the time of creation; 3) by equating inaction of creators with permission for the archive to retain the material, the integrity of the archive's content is put at risk, as any material (e.g. an image within a blog post) may be later requested to be removed; 4) the selection of a designated community is also largely washed over in the web context: in the case of blogs, there is no clear long-term readership, as evidenced by the constantly fluctuating statistics available through search services such as Technorati[13]; 5) the long term deterioration of integrity (through missing objects and lapsed URLs) will result in semantic gaps in the knowledge base; 6) the notion of an intellectual entity is also blurred (e.g. see [2]): new blog posts are added to blogs periodically, previously submitted posts and comments are modified, deleted, and rearranged, changing rights, content and semantics.

As a solution for point 6), some have introduced the notion of archiving versions at varying times as independent intellectual entities. Others have tried to break down the blog into smaller intellectual entities (e.g. posts, comments, embedded objects). This approach could lead to: 1) an unmanageable increase in data storage, 2) many instances of semantically incomplete objects (posts often make sense only in the context of other posts, and even more so for comments and embedded images), and, 3) millions of objects with minor differences between them.

TOWARD PRESERVING CHANGE

We emphasise the *predominance of change* as a core characteristic of today's digital information environment. Change has, of course, always been an integral part of digital information. As we access, save, and transmit information, we cause change and deterioration. To ensure that information does not change from its original state has become essentially impossible [5]. The core purpose of preservation is, not to capture the illusory static steps in between changes, but to ensure that we capture the change itself, and, preserve how changes might propagate other changes. The time dimension in the preservation of the webpages has already been recognised[14,15] but the current paradigm is to understand change as the time-stamped objects in selected states. Ontologies have been proposed to capture events and relations between objects (e.g. see Event Ontology[16] used to represent musical performance; ABC Ontology proposed for preservation [3]). While ontologies provide a step in the right direction, they still describe transitions of object states. Our contention is that objects are symptoms of dynamic processes generated by the medium through which they are broadcast. These need to be captured as recurring patterns within medium dependent event windows that go beyond object boundaries (Figure 3.1).

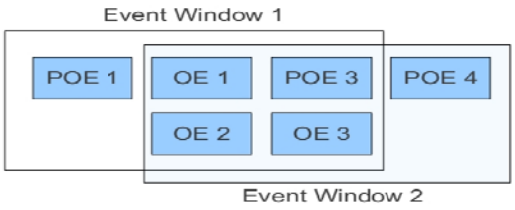


Figure 3.1. Windows of size 3 surrounding physical object events (POE) and other events (OE).

To quote Marshall McLuhan: "the medium is the message" [4]. There is semantics beyond the content of a message: the emergence of so many different channels of communication (e.g. blogs, twitter and facebook) may be a testament to the part that the medium plays in conveying meaning and purpose.

ACKNOWLEDGMENTS

The research leading to the discussion in this paper was conducted as part of the BlogForever project funded by the European Union's Seventh Framework Programme (FP7-ICT-2009-6) under grant agreement n° 269963.

REFERENCES

[1] CCSDS (2002) "Reference Model for an Open Archival Information System (OAIS)", *CCSDS 650.0-B-1* (2002): http://public.ccsds.org/publications/archive/650x0b1.pdf

[2] Hank, C., Choemprayong. S., and Sheble, L. (2007) "Blogger perceptions on digital preservation." In Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL '07). ACM, New York, NY, USA. http://doi.acm.org/10.1145/1255175.1255276

[3] Lagoze, C. and Hunter, J. (2002) "The ABC Ontology and Model", Journal of Digital Information, Vol 2, No 2, http://journals.tdl.org/jodi/article/viewArticle/44

[4] McLuhan, M. (1964) Understanding Media. Routledge, London.

[5] Montague, L., Nicchiarelli, E., Matthezing, H., Kummer, R., Puhl, J., & Roberts, B. (2010b) "The concept of significant

[13] http://technorati.com/

[14] Compare with approaches at http://www.mementoweb.org/

[15] Denev et al. (2011) The SHARC framework for data quality in web archiving. *VLDB Journal*, 20(2):183—207.

[16] http://motools.sourceforge.net/event/event.html

properties", The National Archives, UK & The Austrian National Library

[6] PREMIS Editorial Committee (2011) PREMIS Data Dictionary for Preservation Metadata Version 2.1, *PREMIS Editorial Committee*: http://www.loc.gov/standards/premis/v2/premis-2-1.pdf

[7] Roche, Xavier (2006) "Copying Websites", *Web Archiving*, Julien Masanes (ed.) Database Management and Information Retrieval, Springer, Pp 93-114: http://www.springer.com/computer/database+management+%26+information+retrieval/book/978-3-540-23338-1

[8] Wilson, Carl (2008) "Planets Interoperability Framework Guidelines for Service Wrapping", Planets Project, England: http://sherpa.bl.uk/113/01/Planets_IF6-D2_GuidelinesForServiceWrapping_Ext.pdf

Transformation Rules for Model Migration in Relational Database Preservation

Arif Ur Rahman
Faculdade de Engenharia, Universidade do Porto
INESC Porto
badwanpk@fe.up.pt

Gabriel David
DEI - Faculdade de Engenharia, Universidade do Porto
INESC Porto
gtd@fe.up.pt

Cristina Ribeiro
DEI - Faculdade de Engenharia, Universidade do Porto
INESC Porto
mcr@fe.up.pt

ABSTRACT

Digital preservation is about memory and giving easy access to it. If the digital object is a relational database the requirements of normalization may make it hard to access and understand. In order to deal with this problem we have proposed the DBPreserve approach to transform a relational database to a dimensional model as part of the preservation process, making the preserved information more explicit and easier to access. The paper presents a set of transformation rules to deal with aspects of the migration process such as the identification of the fact tables corresponding to the main organizational processes and the choice of the set of relevant dimensions. The rules help to keep the traceability of the migration process and to preserve integrity and authenticity. The rules were implemented in a case study which involved a human resources information system.

Keywords

Database preservation, database transformation rules

1. MODEL MIGRATION

A relational database incorporated in real information systems normally has a complex structure, integrity constraints, triggers, functions stored procedures and applications developed in high-level language. This makes preserving and using the information in the future difficult. The DBPreserve approach proposes a solution for preserving relational database systems for the future [1]. A relational database is migrated to a dimensional model to make it simple to understand as well as easy to access. In the process of migration, data transformations are needed as changes may occur in the structure and representation of data.

Dimensional modeling is a logical design technique that seeks to present the data in a standard framework which is intuitive, allows for high-performance access and is resilient to change. The strengths of dimensional modeling make it a better choice for long term preservation and access of information.

In the process of migration the information embedded in code is calculated and explicitly stored. This makes the dimensional model independent of the DBMS details and application logic. In the sequel we propose a set of transformation rules which help to effectively carry-out the migration process.

iPRES2011, Nov. 1–4, 2011, Singapore.

2. TRANSFORMATION RULES

Implementing the transformation rules corresponds to concrete extraction, transformation and loading processes where data is selected, cleaned, formatted accordingly, checked for referential integrity against dimensions and transferred. The rules can be grouped into categories. They include:

1. **Generic Table Information**
   (a) Description of each table is prepared including table level information such as name of the table, number of rows, number of columns and a short description. Furthermore, the description also contains column level information such as column names, number of nulls, data type, a short description, distinct values, minimum and maximum values in the column.
   (b) Tables with no data are analyzed and if there is no need to keep them, they may be ignored. In each table there may be some columns which are empty for all the rows. They may also be ignored.
   (c) Sometimes snapshots of tables are taken on a specific date and kept in the database. Also, there may be tables which store data as a preliminary step for all or part of it to be included in a database. If tables are found to be of this nature, they may also be ignored.
2. **Keys**
   (a) The primary keys of tables need to be recorded. Primary keys should be known for the tables which are potential candidates for dimensions though it is not necessary for all tables.
   (b) Foreign keys in tables are also recorded. Situations may arise where there maybe orphan child

records in tables. Orphan child records need special attention in the migration process in order not to lose them.

3. **Processes, Facts and Dimensions**

   (a) Tables are clustered considering their foreign keys and under the broader context of the relevant processes in the organization. The organizational processes about which the system stores data are found out.

   For each process the set of participating tables is listed. Furthermore, in each set of tables there is normally a central table which is identified. Usually, it has no incoming references but it references other tables. The central tables have the real world facts recorded in the organizational processes. They may be candidates to be loaded into fact tables and become the centers of stars.

   (b) Dimensions involved in each organizational process are identified. Analyzing each cluster of tables and having identified which tables are likely to be the sources for the fact tables in the future stars, the remaining tables are candidates to be the source of dimensions. However, their identification is more accurate if it is guided by the knowledge of the organizational processes and of their main entities. Once all the dimensions are identified a bus-matrix is constructed.

   (c) The migrated model should be made of simple stars, to be easy to query. One technique to achieve this is to de-normalize the dimensions, including simple or multiple hierarchies in each one of them. This corresponds to merging tables in the original model.

   (d) There may be situations where a set of tables in the operational system may need to be joined for constructing a dimension and one of them is a lookup table with more records than actually used by the lower level in the hierarchy. In such situations a snowflake schema is constructed to keep the higher level rows.

4. **Nulls**

   (a) In the process of migration if nulls need to be replaced, they should be replaced with a value which has no meaning in the domain.

5. **Code**

   (a) In a database system the application program typically has forms for adding new data, displaying the data already in the system and generating reports. Screen shots of the forms are taken and preserved.

   (b) **Short description of algorithms (code)**
   If there are functions, procedures or code in any form to derive information from the data stored in a database, they are executed and the results are explicitly stored. Furthermore, a description of each piece of code explaining the code and the information it produces is written and kept in the preserved database.

The mappings between the original and the migrated models, which is the base of the ETL process, must be kept as preservation metadata to document the whole process, remain as evidence of the data origin, and facilitate any verification procedure.

3. CASE STUDY

The transformation rules presented in Section 2 were used in a case study. It involved the human resources information system of a higher education institution. The database stores all the information required by the institution to manage the information on teachers and the administrative staff.

In the case study a mapping between the original and the migrated models, which are the base of the ETL process was developed. This mapping was kept as preservation metadata to document the whole process, remain as evidence of the data origin, and facilitate any verification procedure. The information gathered according to rule number 1 helps in performing a completeness check on the data. The recording of keys in rule number 2 makes clustering tables easy. For identifying the organizational processes about which the system stores data, an analysis of the application software used to interact with the database was very helpful. `Contract` is the main organizational process about which the system stores data. Each time a new employee is hired, promoted, assigned extra duties or retired, the data is recorded by this process. In the dimensional model the fact table stores information like the hire date, the contract renewal or expiry date, monthly salary, duration and so on. The fact table is surrounded by dimensions which store the biographical data, the cadre, the unit where an employee works and a date dimension.

In the migration process null values needed to be replaced in columns of type date and character. In the date column the nulls were replaced by 01 Jan 0001 and 31 Dec 9999 depending on the situation and in character type columns the nulls were replaced by 'Unknown'.

The users of the system were involved in preserving the database which helped in carrying-out the task. The migration process also gave the opportunity to detect any missing information and wrong information. It was notified to the users and was corrected.

Although sharing with traditional data warehouse systems many intuitions and techniques, the ultimate goal of preserving a database is very different from the usual goal of building a decision support system. This has some consequences in the nature of the fact tables, which often lack clear measures or the measures included are just secondary elements.

4. REFERENCES

[1] A. U. Rahman, G. David, and C. Ribeiro. Model migration approach for database preservation. In *International Conference on Asian Digital Libraries (ICADL)*, volume 6102, pages 81–90, Springer-Verlag Berlin Heidelberg, 2010.

Considerations for High Throughput Digital Preservation

Jason Pierson
FamilySearch
1221 N Research Way
Orem,
UT, 84097
(1) 801 240 5836
jpierson@familysearch.org

Mark Evans
Tessella Inc
51 Monroe St #702
Rockville,
MD, USA 20850
(1) 240 403 7502
mark.evans@tessella.com

Dr James Carr
Dr Robert Sharpe
Tessella plc
26 The Quadrant,
Abingdon Science Park
Abingdon, Oxfordshire, UK
(44) 1235 555511
james.carr@tessella.com
robert.sharpe@tessella.com

ABSTRACT

In partnership with Tessella, FamilySearch is developing an automated approach to large scale digitization, ingest and long-term preservation of electronic content. The set of proposed processes and underlying architecture must support required ingest rates in excess of 20Tb a day.

Significant effort has been placed on examining the preservation architecture and processes for potential bottlenecks. Digital preservation requires computational intensive capabilities to provide functionality such as fixity checking, format identification and characterization of content. When operating at very large scale there is also a real need for a large network bandwidth and high speed storage systems.

By minimizing the need for human interaction and employing software parallelization our initial findings indicate that the primary bottleneck is not processor bound, but is directly associated with the movement of digital files into and within the application. In short the scalability problem is really a system engineering problem and not necessarily an issue for digital preservation per se.

Keywords

Digital Preservation, Digital Archiving, Scalability, Automation

1. INTRODUCTION

Since the 1930's FamilySearch have been actively involved in capturing images of records that have genealogical significance from all over the world. Up until recently content was captured using film cameras and preserved mostly on microfilm. To date FamilySearch have amassed more than 3.3 million rolls of microfilm in their records vault in the canyons above Salt Lake City UT, USA

In recent years a transition has been made to capturing the content in digital form. Capture rates are currently in excess of 130 million images a year[1]. It is anticipated that by the year 2020 this rate will have doubled. FamilySearch are also in the midst of an aggressive digitization effort with the intent to create digital copies of all the images from all existing microfilm rolls. The combined volume of content from both sources is expected to be in the order of 9Pb a year.

The volume of digital content and the required rate of ingest place large demands on both the processes involved in digital curation and preservation, and the design of infrastructure to support those processes. In order to keep pace with the desired ingest rate a fully automated process is required.

Over the past several years FamilySearch have developed a digital preservation pipeline for the lifecycle management of their digital content. The architecture and design has been driven by the need for both automation and scalability from the very beginning. A recent initiative has been to focus on the long term preservation aspects of the digital content, in particular to look beyond bit level preservation capabilities. This new initiative will use a phased approach with the Tessella Safety Deposit Box (SDB) platform initially running in parallel with the preservation component of the existing preservation pipeline.

2. DIGITAL PIPELINE WORKFLOW

The proposed digital pipeline for the ingestion and storage of content consists of the following processes:

Content acquisition: Digital content is primarily acquired as uncompressed Tiff and raw format from both microfilm scanning and digital camera capture.

Content preparation: Following acquisition, each image is de-skewed, cropped and enhanced using a suite of tools. JPEG2000 and JPEG derivatives for each image are created for preservation and dissemination purposes, and technical metadata to support long term preservation is extracted.

Ingest: The preservation copies and associated metadata are packaged into a SIP and ingested into the SDB platform. During ingest the following operations are performed:

- **Fixity checking –** All content files within a SIP are checked
- **Content and Metadata Integrity –** Checks are made to ensure the right content has been delivered and the metadata is correct and valid
- **Characterization** – The capture of technology-dependent properties of the content files, and the capture of technology-independent significant properties of the information object. This includes **format identification, format validation** and **property extraction.**

Storage: AIP's are stored logically with the metadata and relationships stored in a database and the content stored on a

iPRES2011, Nov. 1–4, 2011, Singapore.

separate file system. In the initial deployment this will be to network attached disk, but in subsequent phases tape will be the primary storage medium. A background process performs periodic fixity checking of all content files once they are persisted.

3. SCALABILITY CONSIDERATIONS

The projected ingest rates for combined microfilm scans and digital camera images based on the estimated volume of microfilm digitization and digital capture are illustrated in table 1.

Table 1: FamilySearch projected ingest rates

Year	Objects per second	MB per second
2011	25	346
2013	42	598
2014	58	776

A major activity to date has been to perform a comprehensive activity of scalability testing to determine the optimum configuration of SDB in order to meet the ingest requirements and identify potential hardware and software bottlenecks.

3.1 Testing configuration

Early testing indicated that a single server approach for SDB would not be sufficient to achieve the projected ingest rates, despite executing multiple ingest workflows in parallel. The main limiting factors in this case were a combination of CPU activity and i/o rates. As a result the SDB architecture has been modified to enable a clustered environment that can utilize a shared storage infrastructure. The actual test environment consisted of two SDB instances (Job Queue servers) connected to a large scale GPFS file system.

The test data was representative of that normally ingested by FamilySearch. Each SIP contained 50-50 mix of JPEG2000 (~10MB each) and XML metadata files (~5K each)

3.2 Variable thread count

A series of tests were conducted where the number of threads (concurrent ingest workflows) on each Job Queue server was steadily increased, until a degradation in throughput was observed.

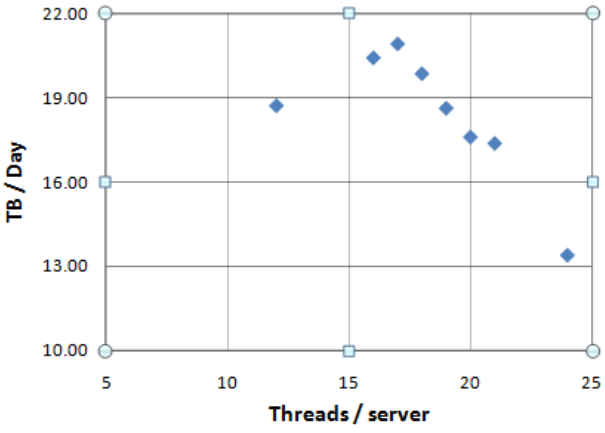


Figure 1: Effects of varying the number of concurrent workflows

For each test a total of 0.5TB was ingested, by using a SIP size of around 4GB and ingesting 120 SIPs.

For the hardware used in this deployment we found running 17 concurrent workflow steps on each server gave the maximum throughput. A subsequent increase in thread count causes a sharp decline in performance, due to the ever increasing demands on the storage system. This is illustrated in figure 1

3.3 Varying SIP size

A second series of tests were conducted to examine the effects of SIP size on the throughput. This was achieved by changing the number of files defined within a submission package, and hence keeping an individual file size constant. As the submission information package grows, the number of ingest workflows required to ingest a fixed amount of content reduces. Since we treat each submission package as an atomic unit, overall reliability may be increased by using larger numbers of small SIP's over a few large SIPs.

The table below shows that the extra cost overhead of running more small SIPs is not significant since the ingest rate is only reduced by around 5% when using 1GB SIPs over 4GB SIPs.

Table 3 Ingest rate results based on varying the SIP size.

	SIP Size				
	4GB	**2GB**	**1GB**	**0.5GB**	**0.25GB**
# Files per SIP	930	466	234	118	60
# SIPS	240	484	925	1749	3701
Time / Sec	8,555	8,739	8,773	9,036	10,408
Files /sec	26.09	25.81	24.67	22.84	21.33
% change	0	+1.07	+5.44	+12.45	+18.24

4. CONCLUSIONS

Our work to date has demonstrated that it is possible to implement a very large scale digital preservation solution. By using a parallelization approach there appears to be no practical limit on the software stack on the rate of ingest; it is more a restriction of the underlying hardware. There are three main areas of future investigations.

- Continuous improvements in the ingest process..For example the performance of format identification may be improved if only a limited number of file formats are being managed.
- Testing in other functional areas such a migration services, and periodic fixity checking
- Consideration of tape storage systems

5. ACKNOWLEDGEMENTS

The authors would like to thank Tom Creighton, Randal Stokes and Steve Lowry from FamilySearch and Alan Gairey from Tessella for their support, dedication and enthusiasm through this collaborative initiative.

6. REFERENCES

[1] Creighton, T; Evans, M; “Digital Object Curation at Scale”, Proceedings of Archiving 2011, Salt lake City UT, May 2011.

How Clean is Your Software? The Role of Software Validation in Digital Preservation Research Projects

Leo Konstantelos
HATII, University of Glasgow
11 University Gardens
Glasgow G12 8QH, UK
+44 141 330 7133
leo.konstantelos@glasgow.ac.uk

Perla Innocenti
HATII, University of Glasgow
11 University Gardens
Glasgow G12 8QH, UK
+44 141 330 4453
perla.innocenti@glasgow.ac.uk

Seamus Ross
Faculty of Information, Univ. of Toronto
140 St. Georg Street Toronto
Ontario M5S 3G6, Canada
00 1 416 978 3202
seamus.ross@utoronto.ca

ABSTRACT
The emphasis during the last three EC Framework Programs on Digital Preservation (DP) has enhanced our understanding of the boundaries of the problem, produced new methods, and supported the construction of preservation tools. In the case of methods and tools, measuring their suitability has emerged as a focus of research. What evidence does the DP community have that the digital preservation tools perform the functions they are supposed to and that they align with original specifications? This paper summarizes the efforts of the Sustaining Heritage Access through Multivalent ArchiviNg (SHAMAN) project to create and implement software validation as a means to gauge the fitness-for-purpose of software outputs. The paper concludes by demonstrating the applicability of the validation approach to preservation projects in general.

Categories and Subject Descriptors
D.2.4 [**Software Engineering**]: Software/Program Verification – *validation*

General Terms
Measurement, Performance, Reliability, Verification.

Keywords
digital preservation, software validation, evaluation, SHAMAN.

1. INTRODUCTION
In the last five years or so, the Digital Preservation (DP) domain has witnessed an unequivocal growth in investment, accompanied by an equally growing number of software applications that encapsulate the aims and objectives of DP research projects. Evidence on the existence, suitability, merit and functionality of these programmatic tools has been documented as part of the reports delivered under the banner of an array of DP projects funded by the European Commission.

Through this evidence, the role and significance of DP-specific software becomes clear to both funders and the community of users. Test implementations, demonstration versions and free open-source software outputs create a matrix of available options. Although the existence of these software tools in the public domain testifies the achievements of DP projects, there is an underlying layer, whose importance has – thus far – failed to emerge. What evidence does the DP community have that the digital preservation tools perform the functions they are supposed to, based on user requirements and the assertion of Descriptions of Work accompanying funded DP projects? How often is it that the alignment of software developed under these projects with original specifications are formally documented and made publicly available? This paper brings attention to the use of software validation [3, 6] as a means to significantly boost the trustworthiness of DP software developed under EC-funded projects. At present, there is a distinct lack in the DP domain of not only documentation of software validation, but of implementation of the very process itself. In a field which still strives to justify funding and investment, software validation can contribute to advocating the quality of software outputs from research projects for digital preservation, and support the claims for their suitability to actively aid in safeguarding the digital production of the 21st century. In recognition of this situation, this paper summarizes the efforts of the Sustaining Heritage Access through Multivalent ArchiviNg (SHAMAN) Integrated project to create and implement software validation for its software outputs.

iPRES2011, Nov. 1–4, 2011, Singapore.

2. THE SHAMAN PROJECT
The SHAMAN Integrated project[1] has received funding from the 7th Framework Program for research and technological development of the European Commission to conduct research in the long-term preservation of digital heritage. SHAMAN centers on three distinct domains of focus, namely Memory Institutions, Industrial Design & Engineering and e-Science. Each domain has been commissioned to generate functional research prototypes, which "exhibit, test and validate the principles, functionality, viability and usefulness of the SHAMAN solutions" [5]. The combined results from the ISPs in all three domains of focus form the basis for achieving the SHAMAN high-level goal, that is to develop a next generation Digital Preservation (DP) Framework. The SHAMAN DP Framework is meant to provide a solution to

[1] http://www.shaman-ip.eu

the challenges posed by technological evolution and obsolescence through sustainable, scalable services for the preservation of complex objects and their relationships. SHAMAN has developed demonstrator applications[2] as exemplar cases that proclaim the operations and benefits of the prototypes and form the primary software outputs of the project.

3. SOFTWARE VALIDATION METHODOLOGY

The project partners in the Humanities Advanced Technology & Information Institute (HATII) at the University of Glasgow were appointed with the task to build and implement an instrument to assess software development. Guided by the specifications in the IEEE V&V Standard [3] the HATII team devised a validation methodology to complement the design and development of the SHAMAN demonstrators. The methodology was based on the IEEE Standard in order to generate a custom validation plan for the SHAMAN demonstrators. A set of validation instruments was selected that align with the specifications of the IEEE Standard. A comprehensive presentation and explication of these tools is included in the SHAMAN Report on Demonstration and Evaluation Activity in the Domain of Memory Institutions [2].

4. SOFTWARE VALIDATION RESULTS

The information collected through the metrics identified in the validation methodology provided valuable feedback and corroborated the alignment of project specifications to the development and implementation of demonstration software within the domain of memory institutions. At the conclusive stage of the demonstrators' development, the involved programmers and software developers were invited to validate the level of achievement with respect to the functional requirements identified through a software validation process. This was achieved through a dedicated online submission system, which collected feedback for a period of one month.

The aggregate results show that 87% of the total functional requirements across all ISP1-related Use cases have been completely or partially achieved. The requirements that were not achieved are straightforwardly implementable and are addressed in demonstrators for the subsequent ISPs. The level of overall achievement corroborates that the software developed at this phase of the project satisfies the original expectations and aligns with both the SHAMAN Description of Work and the needs to demonstrate the SHAMAN DP Framework to its community of users through exemplar software implementations. The results from the validation process show that the SHAMAN ISP1 Demonstrators have satisfied and justified the reasons for the development, covering at the same time fundamental project needs to showcase a real-life example of implementing the SHAMAN DP Framework.

5. CONCLUSIONS

Validating and assessing the achievement of DP projects in terms of developed software tools, in a structured and objective manner, is essential for both ensuring continuing investment in this field and for guaranteeing the good operation of these tools to target communities of users. For software validation to be successful the process must be included in the strategic plans of DP projects at an early stage and consistently continue to be applied until the end of the projects' life-cycle. In this manner, it is possible to exhibit verification and validation evidence that software development directions are not a tangent to user expectation. Software validation – although not a panacea for all software hardship a project might encounter – is a suitable evaluation and assessment method for measuring, quantifying and ascertaining the quality of software produced via research projects and its fitness-for-purpose. This paper concludes with the question it started from, addressed to anyone involved in development of research software for digital preservation: How clean (really) is your software?

[2] The Memory Institution Demonstrator is available via registration at https://shaman-ip.eu/shaman/Demonstrator%20for%20Memory%20Institutions

6. ACKNOWLEDGEMENTS

The research presented in this paper has been developed within the SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg) Integrated Project, co-funded by the European Union FP7 (Grant Agreement No. ICT-216736).

7. REFERENCES

[1] Aitken, B., Innocenti, P., Ross, S., and Konstantelos, L. 2010. User Requirements for a Next Generation Digital Preservation Framework: Analysis and Implementation. In *Proceedings of the Archiving 2010 Conference* (Den Haag, the Netherlands, May 2010). Society for Imaging Science and Technology, pp. 48-52.

[2] Birrell, D., Menzies, K., Maceviciute, E., Wilson, T., Wollschlager, T., Konstantelos, L., Innocenti, P., Riestra, R., Lindh, M. Harrison, J., and Hasan, A. 2010. *Report on Demonstration and Evaluation Activity in the Domain of Memory Institutions*. SHAMAN Technical Report WP14-D14.2. URL= http://shaman-ip.eu/shaman/sites/default/files/SHAMAN%20D14.2_Report%20on%20Demonstration%20and%20Evaluation%20activity%20in%20the%20domain%20on%20MI_0.pdf

[3] IEEE Computer Society. 2005. *IEEE standard for software verification and validation*. New York, N.Y: Institute of Electrical and Electronics Engineers.

[4] Innocenti, P., Aitken, B., Hasan, A., Ludwig, J., Maciuvite, E., Barateiro, J., Antunes, G., Mois, M., Jäschke, G., Pempe, W., Wilson, T., Hundsdoerfer, A., Krandstedt, A., and Ross, S. 2009. *SHAMAN Requirements Analysis Report (public version) and Specification of the SHAMAN Assessment Framework and Protocol*. SHAMAN Technical Report D1.2. URL= http://shaman-ip.eu/shaman/sites/default/files/SHAMAN_D1_2Requirements%20Analysis%20ReportSHAMAN%20Assessment%20Framework.pdf

[5] Innocenti, P., Konstantelos, L., Ross, S., Maceciuvite, E., Wilson, T.D., Ludwig , J. and Pempe, W. 2010. Assessing Digital Preservation Infrastructures: A Framework for Library, Engineering and eScience Organisations. In *Proceedings of the Archiving 2010 Conference* (Den Haag, the Netherlands, May 2010). Society for Imaging Science and Technology, pp. 18-23.

[6] Wallace, D. R., and Fujii, R. U. 1989. Software Verification and Validation: An Overview. *IEEE Software*, 6,3, 10-17.

Long-Term Storage Features of Optical Disks According to Recording Conditions

Kwan-Yong Lee
Center for Information Storage Device,
Yonsei University
YERC 332D, 134 Shinchon-dong,
Seodaemoon-Ku, Seoul, Korea
+82-2-2123-4677
gladiater9@yonsei.ac.kr

Won-Ik Cho
Toshiba Samsung Storage
Technology Korea (TSST)
14rdFloor, 102Bldn, Digital Empire2,
486-1, Sin-dong, Yeongtong-gu,
Suwon-si, Gyeonggi-do, Korea
+82-31-8006-6363
wonik.cho@samsung.com

Young-Joo Kim*
Center for Information Storage Device
Dept. of Mechanical Engineering
Yonsei University
YERC 335B, 134 Shinchon-dong,
Seodaemoon-Ku, Seoul, Korea
+82-2-2123-6852
yjkim40@yonsei.ac.kr

ABSTRACT
Optical disks are widely used in libraries and archives as digital data media due to their long-term storage stability. Though archive-grade optical disks are already available on the market, there is relative less focus on the reliable recording conditions. Commercial DVD-R media were recorded at various recording speeds with the maximum speed of 16X, and tested at acceleration aging conditions. Through the evaluation of long-term storage features by the data stability test, a higher recording speed over 12x resulted in better long-term storage stability.

Keywords
Long-term data storage, Optical disks, Archival application, Recording condition, Data stability

1. INTRODUCTION
Optical disks have considerable potential to provide reliable long-term data storage since it has little influence from the environmental electromagnetic fields and physical shocks. In addition, the contactless reading process of optical data storage promises undamaged multiple information playbacks [1]. For these reasons, optical disk is considered a long-term storage system for library and archival applications. Though there are few reports related to the influence of recording conditions on storage reliability, many users have confirmed the advantages of optical data storage. Thus, we studied the effect of recording speed on data stability by evaluating the quality of recorded data through acceleration conditions.

2. EXPERIMENTAL PROCEDURE
A different kind of commercial DVD writer and DVD-R media (1-16x) were prepared for the recording experiment. Current optical disk drives provide a range of recording speeds from 1x 8x, to 16x. The recording speed is controlled by two different modes such as a constant linear velocity (CLV) and constant angular velocity (CAV). The CLV mode provides a constant linear speed through all recording tracks by controlling the speed of spindle motor. On the other hand, the CAV mode fixes the motor speed with the same angular velocity, then the linear speed in the recording tracks increases from the inner tracks to outer tracks. The CAV mode provides shorter recording time than the CLV mode since the rotation speed of spindle motor at the CAV mode is kept as the highest value. In this study, four different recording conditions were selected as shown in Fig. 1.

iPRES2011, Nov. 1–4, 2011, Singapore.

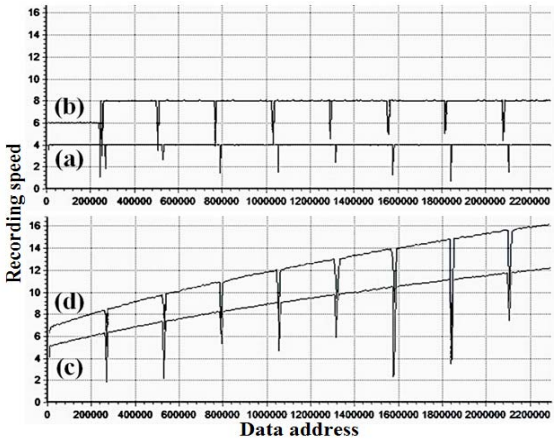


Figure 1. Various recording conditions at (1)4x: CLV, (2)8x: zone-CLV, (3)12x: low speed CAV, (4)16x: high speed CAV.

After being recorded at different recording conditions, the DVD-R media were stored for 1000 hours (250 hours/cycle, 4 cycles) at acceleration conditions to estimate the archival lifetime of optical media followed by the ISO/ICE 10995 [2]. In this study, the most severe condition of 85ºC/85%RH was selected to understand the effect of recording speed. The data stability of test disks was measured by measuring the parity inner code error sum 8 value (PI8) after every acceleration test cycle. The PI8 value means the total number of parity inner errors in any 8 consecutive error correction code (ECC) blocks. The value of 280 for PI8 is considered as a limit of stability without the error correction [2]. The PI8 value was measured in all tracks using an optical disk analyzing software with a special emphasis on the maximum PI8 which is considered as the most important value for stable playback. The experimental procedure is shown in Fig. 2.

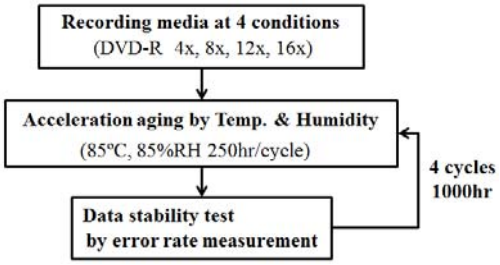


Figure 2. Experimental procedure for the stability test.

3. RESULTS & DISSCUSSIONS

3.1 Initial state after recording

The playback stability results of DVD-R media after being just recorded are shown in Fig. 3. The PI8 evaluation results show a stable value less than 100 with all media as well as at whole tracks. The distribution of PI 8 value is roughly uniform through the track positions. Thus, the data stability with a range of recording speeds shows an almost same value at the initial recording state.

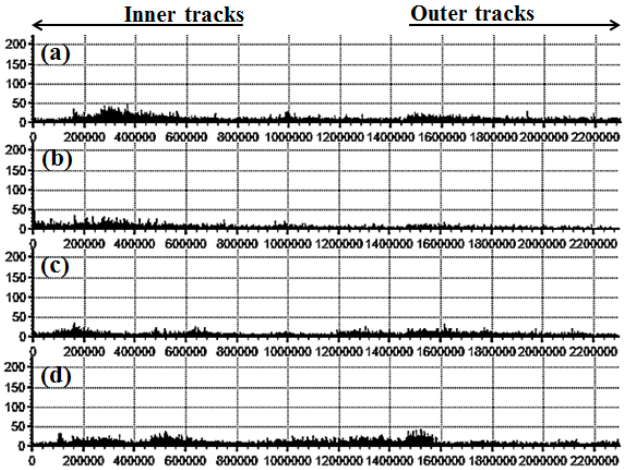


Figure 3. PI error sum 8 evaluation results at the initial state with different recording speeds of (a)4x, (b)8x, (c)12x, (d)16x.

3.2 Acceleration test results

The PI8 values were measured at a series of aging time with a 250-hour interval for the comparison of data stability as shown in Figure 4. Each PI8 value means the arithmetic average PI sum 8 from 10 media. From this graph, we can confirm that the media recorded at 16x speed shows more stable result than that of other recording speeds. It means that long-term storage features can be improved by increasing the recording speed to the maximum value since the recording conditions might be optimized for the higher speed, including a write strategy.

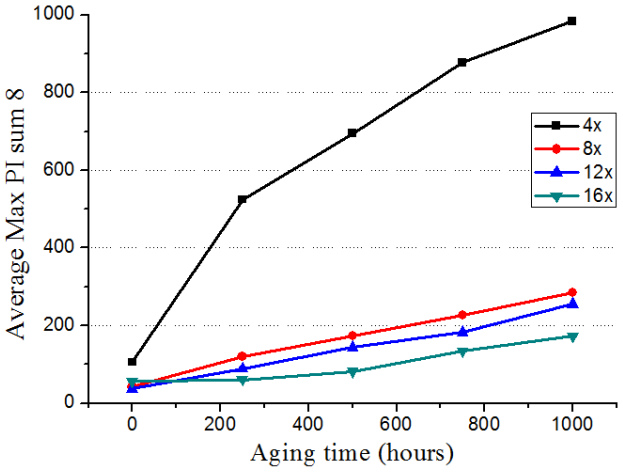


Figure 4. The average PI sum 8 value of DVD-R media as a function of recording speed after the acceleration aging test at the sever condition of 85ºC/85%RH.

3.3 Data stability test results

The PI8 values as a function of track positions were also observed as shown in Figure 5. The data stability of outer tracks increases at the higher recording speed such as Fig. 5(d). Since the higher speed (12x-16x) was realized at the outer track, more stability was observed at the outer tracks for the media recorded at the16x speed. As expected, the media recorded at lower speed showed less stability through whole tracks as shown in Fig. 5(a).

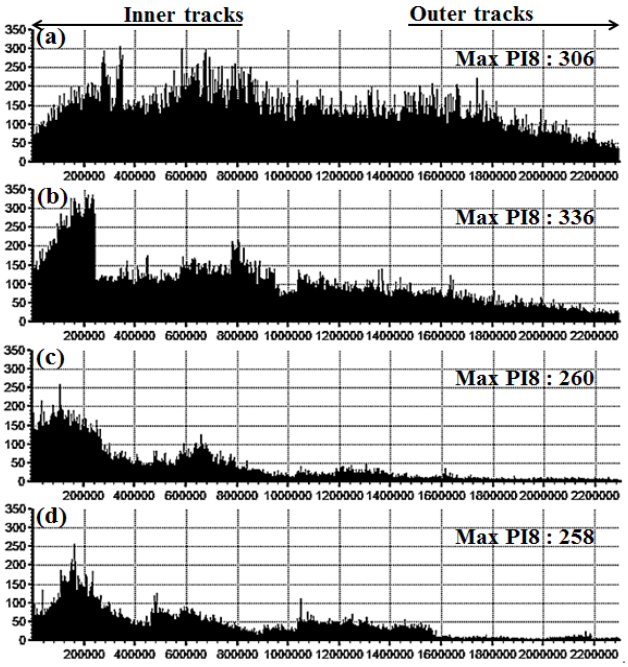


Figure 5. The PI8 evaluation results for the media recorded at (a)4x, (b)8x, (c)12x, (d)16x after acceleration aging test.

4. CONCLUSIONS

The data stability of optical disks was evaluated by the acceleration aging test according to various recording speeds. Though it is generally known that lower recording speed such as 4x provides a reliable data stability, we understood that the higher

recording speed (16x in this study) showed more stable long-term storage feature. In addition, the outer track recorded at higher speed (12x-16x) shows better data stability compared to the inner track recorded at a lower speed. Thus, if the recording parameters in the drive were optimized for a higher speed such as 16x, the higher recording speed might be recommended for long-term storage.

5. ACKNOWLEDGMENTS

This study was executed as a part of the Research and Development Project of the Archives Preservation Technology hosted and supported by the National Archives of Korea, the Ministry of Public Administration and Security, for which we would like to extend our sincere gratitude.

6. REFERENCES

[1] Chandru J., Basil M., and Michele Y. 2003. Longevity of CD Media: Research at the Library of Congress. http://www.loc.gov/preservation/resources/rt/studyofCDlongevity.pdf

[2] ISO/IEC 10995. 2008. Test method for the estimation of the archival lifetime of optical media. Geneva, Switzerland, 2008-04-15.

Curation and Preservation of Research Data in Germany: A survey across different academic disciplines

Achim Osswald
Cologne University of Applied Sciences
Institute of Information Science
50968 Cologne, Germany
achim.osswald@fh-koeln.de

Heike Neuroth
Goettingen State and University Library
37073 Goettingen, Germany
neuroth@sub.uni-goettingen.de

Stefan Strathmann
Goettingen State and University Library
37073 Goettingen, Germany
strathmann@sub.uni-goettingen.de

ABSTRACT
This paper gives an overview of the design and purpose of a survey on the curation of research data in Germany. Eleven disciplines including, among others, the humanities, social sciences, and medicine are addressed. Issues and preliminary findings are summarized. At iPRES2011 findings of this survey will be presented to an international audience for the first time.

Categories and Subject Descriptors
A.1 INTRODUCTORY AND SURVEY

Keywords
Digital curation of research data, metadata, cooperative structures, research archives, cost and funding, training, perspectives and visions, scientific communities, survey, Germany.

1. INTRODUCTION
In the last few years the issue of curation of research data has become a topic of enhanced interest in scientific communities. Awareness of long-term availability, re-usability and integrity of research data has been stimulated by international reports (e.g. by the High Level Expert Group on Scientific Data (October 2010), on behalf of the European Commission) [1]. But even now there is no clear understanding of how to deal with research data curation. Concepts suggested range from attempting to find one promising approach that works for all disciplines to developing specific approaches for every single discipline.

For some disciplines, like astronomy or climate research, international cooperation among research institutions in several countries has already been established (see for example the World Data Systems). Others have not yet begun to address the problem.

On the national level, supporting or stimulating activities like implementing a data management plan have been realized in some countries, e.g. by the NSF (National Science Foundation, USA), where the submission of such a plan is required, or the DFG (German Research Foundation), where a data management plan is strongly recommended. The ANDS (Australian National Data Service) has gone a step further and initiated the Australian Research Data Commons (ARDC).

iPRES2011, Nov. 1–4, 2011, Singapore.

2. BASELINE STUDY: RESEARCH DATA IN GERMANY
In Germany libraries, archives, museums and leading experts in the field of digital curation and digital preservation work together in *nestor*, the German competence network for digital preservation. Their objective is to ensure the long-term preservation and accessibility of digital resources [2]. In 2006 a group of experts published the first edition of "nestor Handbuch. Eine kleine Enzyklopädie der digitalen Langzeitarchivierung" (nestor handbook. A small encyclopedia of digital preservation) [3] which is a comprehensive state-of-the-art documentation on digital curation and digital preservation in German language. The special issues of curation and preservation of digital research data are a topic of some small chapters only because activities in most disciplines have been in a very early state of development. The editors like to complement the encyclopedia with a survey about the curation of digital research data.

Regarding the situation in Germany, there is no clear picture of the methods that different academic disciplines use to preserve and curate their research data. There is a need to address the issue with a baseline study. This will give more stable data to scientists, service infrastructure experts and politicians to foster strategic concepts for digital curation and preservation in and between the disciplines.

To broaden and promote access to researchers and disciplines the editors started a cooperation with D-Grid GmbH [4], a non-profit Development and Operating Company founded by the German Ministry of Education and Research (BMBF) in 2008. D-Grid has the goal to ensure efficient collaboration and cooperation between different projects in the field of a sustainable grid infrastructure in Germany.

2.1 Study design
With the support of scientists in the addressed disciplines, the authors and editors conducted a detailed survey across eleven disciplines including the humanities, social sciences, psycholinguistics, pedagogics, classical studies, geoscience, climate research, biodiversity, particle physics, astronomy and medicine. These disciplines have been selected because the type of research data relevant to these fields cover nearly all types of research data. Moreover, it seems that these disciplines have

already started with digital curation of data or have more experience in this field. Therefore, they can be seen as more or less representative for the situation in Germany and will help to get a deeper insight into digital curation and preservation of research data.

Well known and accepted experts from the different disciplines could be gained to serve as authors and contributors to the survey.

A workshop with scientists and scholars in these disciplines in March 2011 showed that there are several issues to be addressed in common. Nevertheless concepts and solutions addressed vary from discipline to discipline. Currently, the results of the survey and the workshop are being evaluated, normalized and a detailed report is being prepared.

2.2 Issues addressed

The report will be published at the beginning of 2012 and will address at least the following issues:

- What types of cooperative structures do already exist? How are they stimulated? What makes them successful?
- What are the types and the amount of data relevant for digital curation and preservation activities?
- What kinds of metadata standards are used? Are there international standards etc., relevant to the discipline in focus?
- Are there any research data archives already dealing with the curation of data in this specific academic discipline? If so, how are they organized and financed? How expensive is the initial funding of such archives and what are the operating expenses per year and in the long run?
- What are the perspectives and visions of data curation and preservation in the different scientific communities?

These issues are part of a larger sample of questions deduced by the editors within a preliminary study of the topic. They have been collected in a detailed survey given to the experts in the different disciplines. This enabled a structured and comparable view to the results collected.

2.3 Preliminary findings

Preliminary findings show that the 11 academic disciplines involved have produced a variety of solutions to the issues. Up to this point, the development of infrastructures to assure the quality of data, to archive it and to assure its long-term availability as well as re-usability has been influenced primarily by traditions and independent infrastructures in the different academic disciplines. Issues of multidisciplinarity and international interoperability are reflected in differing degrees. Scientists seem to be aware that there is a need to find a balance between community-driven approaches and standardization, common policies working for all disciplines versus domain-specific isolated solutions. Although not yet published, the survey has gained some attention inside and outside of Germany. While still in its preparation phase, it initiated an exchange of ideas among the different disciplines. The authors expect that the report will not only illustrate the situation of curation of research data in Germany, but will also stimulate a broader discussion among the different disciplines on an international level.

Major findings of this interdisciplinary survey will be presented at iPRES2011.

3. ACKNOWLEDGEMENTS

The authors of this paper like to thank their collaborating editors, the authors of the discipline-related reports, D-Grid and the Federal Ministry of Education and Research of Germany for their contributions and their support to realize the study.

4. REFERENCES

[1] Riding the wave - How Europe can gain from the rising tide of scientific data - Final report of the High Level Expert Group on Scientific Data - October 2010, http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204

[2] See http://www.langzeitarchivierung.de/eng/index.htm.

[3] See the latest edition of Neuroth, Heike; Oßwald, Achim; Scheffel, Regine; Strathmann, Stefan; Jehn, Matthias (Ed.) (2010): *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*: edition 2.3; http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:0008-2010071949

[4] D-Grid GmbH; http://www.d-grid-gmbh.de/index.php?id=1&L=1

Managing Preservation Networks: Issues of Scale for Scientific Research Assets

Esther Conway
STFC
Rutherford Appleton Laboratory
esther.conway@stfc.ac.uk

Simon Lambert
STFC
Rutherford Appleton Laboratory
simon.lambert@stfc.ac.uk

Brian Matthews
STFC
Rutherford Appleton Laboratory
brian.matthews@stfc.ac.uk

Arif Shaon
STFC
Rutherford Appleton Laboratory
arif.shaon@stfc.ac.uk

ABSTRACT
The preservation of science data requires consideration of a wide range of factors from file formats to analysis software. Previous work has reported on the development of Preservation Network Models that capture dependencies at multiple levels and allow reasoning about preservation planning and actions. However this is only one aspect of the development of a trusted preservation environment; there is also a need for quality assurance and relation to explicit policies on data preservation. This paper presents issues of scale for scientific research assets which will be explored further on the SCAPE project.

Keyword
Digital Preservation, Scientific Data, Preservation Network Models

1. Introduction
Preserving scientific research data has become increasingly recognized as necessary for the long term benefit of large scale data collection to be fully realized. However, the complex nature of science data and its dependencies on software, together with the scale of the data holdings involved make this a major challenge.

Given the sheer number of existing data files and the anticipated increase in production rates, scalability of any preservation action has become an important issue for the archive. Scale is absolutely critical in two respects for cost reduction through the re-use of preservation solutions and automation of preservation action. We consider these issues using Preservation Network Models (PNMs), a preservation analysis methodology which was originally developed within the CASPAR project.

2. Preservation Network Models
A PNM is a formal model for conceptualising the relationships between resources within the scenario of a preservation objective."The preservation network model consist of two components: the digital objects and the relationships between them possesing atrribute of (Information, Location ,Physical state) and (Function, Risks and Dependencies, Tolerance, Quality assurance/ testing) repectively

iPRES2011, Nov. 1–4, 2011, Singapore.

3. Preservation Action
Networks are created and evolve through preservation actions which are made on particular relationships or acknowledged dependencies within the network. Types of action are

- Risk acceptance and monitoring
- Software capture and extension through the stack
- Description
- Migration

4. Quality Assurance
The quality assurance of a preservation solution is provided by two mechanisms Trust in or Testing discussed below

4.1 Trust
Trust occurs when the archive appraises a solution as satisfactory for one of the following reasons

- Trust in a custodial organization; when the archive relies on an external organization to maintain the integrity and supply of important information. The accepted reputation of the organization supplies the required assurance.
- Trust in a standardization process; when an archive acquires descriptive information which has produced as a result of a standardization process such as ISO.
- Quality of Sources; this occurs when an external organization supplies an archive with an information object. Trust is based upon the belief that the supplier has delivered a quality preservation solution.

4.2 Testing
When an archive cannot fully trust a solution it must then employ testing to gain necessary assurance. We consider three testing scenarios.

- Passive testing; occurs when a preservation solution is exposed to an active user community with the expectation that they will report any deficiencies.
- Proactive testing; occurs when external experts are invited to test a preservation solution.
- Direct testing; occurs when the archive conducts testing itself.

5. Monitoring the Preservation Environment
No preservation solution is permanent and will always carry risks due to dependencies. Change is required for a number of reasons detailed below

When a preservation solution is longer valid this forces a reevaluation in terms of new information needs of the user community and the funding available to carry out preservation in a sustainable way.

Most changes are due to the realization of risk causing failures of the preservation solution that are within tolerance, partial or critical The explicit statement of technical dependencies within a network can be used to determine the types of things that need be monitored.

- Dependencies on external organizations risk acceptance which by definition inform watch services
- Dependencies on "software capture" strategies require the monitoring of libraries and operating systems
- Dependencies on a descriptive strategy involving community skill, support and resources

In addition to external triggers which invalidate the preservation solution "risk acceptance and monitoring" also has the capacity to support evolution of the scientific asset. It forms one of a number of positive feedback relationships when multiple strategy types are employed.

6. Preservation Action in a Scalable Environment

In this section we e give the illustrative examples from the ISIS GEM powder diffraction instrument. We explore how each of the main types of preservation action are affected by issues of scale for the creation and maintenance of scientific research assets.

6.1 Monitoring Websites

The preservation network uses two different risk acceptance and monitoring strategies with different degrees of re-use. The archived website held by the UK web archiving consortium hold information which should be universally associated with all data files. However, the software which the Mantid website provides access to is not universally applicable. Data files from different beam lines and experiment types require different forms of analysis. Currently Mantid can support the minimal required analysis for approximately 60% of data holdings. The Mantid website needs therefore to be associated with files whose preservation objective requires the type of analysis supported by Mantid. In both cases automation is required to propagate necessary notifications and changes for example new URL's reference points when websites are migrated or failure of the solution through the networks, as thousands of file should be associated with both these externally managed information objects.

6.2 Capture of Mantid Software

As described above this type of strategy involves the acquisition and management of information objects. As with the risk acceptance and monitoring strategy re-use of this solution (network branch) is appropriate for around 60% of data holdings based on their preservation objective.

Again because of the number of files associated with a software capture solution, automated changes to large numbers of networks become desirable. Removal of platform dependent network branches when operating systems become obsolete or extension of the branches to include libraries and emulators is required in order to stabilize the solution. The automated addition of alternates involving new binaries or source code would also be advantageous. These can then be recompiled to work on different operating systems when communities begin using new technologies analysis techniques.

6.3 Description of Analysis Algorithms

The descriptive strategy a scientist to identify, extract and correctly interpret parameters and the relationship between them. A scientist can subsequently carry out a specified type of analysis by applying the described algorithms. The capture of specific algorithms mean the user is restricted to a particular analysis path which is a functional subset of both the software capture and risk acceptance strategies. As a result the degree to which this solution can be reused by different data files is much lower as experiment types have unique analysis requirements. As this type of preservation strategy is technology agnostic the only automation required is the ability to update the analysis path for multiple networks once the old have been deprecated and new algorithms gain community acceptance.

6.4 Conversion of Document formats

The need for automation becomes important when an archive needs to transform a large number of digital objects from one format to another. When the preservation network models are logical rather than physical there are variations in the numbers of actual objects which may require conversion. If we consider the example of an archive making a decision to convert all word documents to PDF. Automation is not necessary for the word documents describing the experimental environment and preparation methods within the instrument website. While the website is logically referenced by thousands of PNM's there exist only a couple of physical copies making manual conversion the most efficient option. However experimental proposals are unique to a discrete set of data files the ability to characterize the relationship and format of a digital object and automate its transformation is critical to the successful management of information on this scale.

7. Policy Formation

Dealing with large volumes of data with differing preservation objectives can place addition pressures on an archive. Based on our previous discussions we suggest areas where an archive may wish to develop policy to manage such large scale data holdings.

- Policy on number and types of dependencies and archive should be exposed to in order to maintain an acceptable risk burden.
- Policy can specify appropriate levels of vigilance and monitoring of the preservation environment.
- Policy can specify what is a trusted organization, institution or standards.
- Policy can mandate levels of testing required for any preservation solution to be deemed acceptable
- Policy can specify how much of the hardware/software environment should be captured or if the solution should be supplemented with source code.
- Policy can recommend the employment of multiple strategy types to lower risk burden and enhance long term usability
- Policy can also stipulate acceptable formats which an archive a can reasonably expect to support and monitor
- Policy can mandate descriptive preservation solution for non standard formats

8. Conclusions

The use of Preservation Network Models provides a basis not only for preservation planning and actions, but also for other preservation-related aspects such as quality assurance or trustworthiness. By conducting an analysis of dependencies, founded on specified preservation objectives, issues such as scalability can also be analyzed. Furthermore there is an interaction with preservation policies: Preservation Network Models can highlight areas where policies should be put in place, and help to guide their formulation. Thus these models are proving an invaluable framework for scientific data preservation at STFC facilities. Further exploration and trialing on part of the ISIS archive within the SCAPE project to fully address the issues of scale discussed in this paper is required.

The Data Management Skills Support Initiative:
Synthesising Postgraduate Training in Research Data Management

Laura Molloy
Humanities Advanced Technology and Information Institute
University of Glasgow
Glasgow, Scotland
(+44) (0)141 330 2793
Laura.Molloy@glasgow.ac.uk

Kellie Snow
Humanities Advanced Technology and Information Institute
University of Glasgow
Glasgow, Scotland
(+44) (0)141 330 8620
Kellie.Snow@glasgow.ac.uk

ABSTRACT
In this paper, we describe the context, methods and findings of the Data Management Skills Support Initiative ('DaMSSI'), which supported the five JISC Research Data Management Training Materials ('RDMTrain') projects of the JISC Managing Research Data programme ('JISCMRD') in developing discipline-focused postgraduate training units in research data management. The Initiative tested the effectiveness of two skills frameworks and produced a comparison and synthesis of training approaches by the RDMTrain projects. DaMSSI also assisted in the production of a number of guidance documents to raise awareness of the importance of data management in career development.

Keywords
Training, education, skills, skills frameworks, research data management, postgraduate, UK, digital curation.

1. INTRODUCTION
The Data Management Skills Support Initiative ('DaMSSI') was co-funded by the JISC Managing Research Data programme and the Research Information Network ('RIN'), in partnership with the Digital Curation Centre, to review, synthesise and augment the training offerings of the JISC Research Data Management Training Materials ('RDMTrain') projects, a strand of the JISC Managing Research Data ('JISCMRD') programme.

2. BACKGROUND
In recent years, significant effort has been put into defining data management roles and responsibilities for those involved in the production of digital research data. The National Science Foundation's 2005 report [1] suggested a number of responsibilities that data authors should recognise, but despite these recommendations being around for some time, there is still little evidence that data management skills are being embedded within UK postgraduate courses. Feedback from attendees at events such as the JISC Innovation Forum 2008 data management skills and capacity session [2] indicates that while some UK university departments are delivering training to their postgraduates, much more needs to be done to embed data management training into all postgraduate programmes. There is also evidence that researchers in UK HEIs are likely to respond favourably to data management support which is presented with a focus relevant to their discipline [3].

iPRES2011, Nov. 1–4, 2011, Singapore.

3. MAPPING AND SYNTHESIS
To improve the provision of research data management practice at postgraduate level, JISC funded the five projects of the RDMTrain strand [4], with the aim of creating a body of discipline-focussed postgraduate training units which could be reused by other institutions to stimulate curriculum change and create a greater awareness of the need for research data management skills training. DaMSSI worked with and supported the RDMTrain projects in a reciprocal relationship.

With the cooperation of the projects, DaMSSI tested the effectiveness of two skills development frameworks, namely the Society of College, National and University Libraries' ('SCONUL') Seven Pillars of Information Literacy model [5], and Vitae's Researcher Development Framework ('RDF') [6] for consistently describing data management skills and skills development paths in UK HEI postgraduate courses.

The Initiative mapped individual course modules from each project to the two frameworks and basic generic data management skills were identified alongside discipline-specific requirements [7]. Along with highlighting issues about the value of the RDF and Seven Pillars models themselves, the mapping of the projects' course outputs to the models suggested that there was consistency in the data management skills required across the disciplines, despite variety in the arrangement of course modules among the projects. Discipline-specific variations through examples and case studies constituted the main ways courses were further customised.

The mapping and feedback from the projects allowed DaMSSI to identify the extent to which the two models appeared useful for describing and supporting data management training. Overall, we found that the models were potentially useful for describing and embedding courses, but at the same time needed to offer clearer

definitions and more detail on information handling and data management. These findings were fed back to SCONUL and Vitae to help in the further refinement of their models, and both organisations have begun to implement changes as a result.

DaMSSI then embarked on subsequent synthesis work to explore the findings from the mappings further, analysing the projects' own conclusions, final course materials and feedback from students to see if there was agreement with what the mapping had suggested. The project found that generic principles applied across all disciplines, but discipline-specific definitions, examples and exercises were seen as beneficial by both the projects and course participants to demonstrate relevance. Courses that successfully balanced the need for discipline-specific detail with keeping training relatively brief and concise showed better delegate retention, and face-to-face sessions were liked by delegates as a way to share experiences. Data management plans (DMPs) which offered discipline-specific interpretation proved particularly popular, yet despite this students were still reluctant to put effort into DMPs unless there was specific requirement from the institution or funders.

The findings were then refined into a set of recommendations (see section 5).

4. DATA MANAGEMENT SKILLS

DaMSSI supported the production of a number of guidance documents to raise awareness of the importance of data management in a variety of careers. The profiles of conservator, social researcher, archaeologist, clinical psychologist and data manager were produced to link in with the disciplines covered by the RDMTrain projects. The profiles have the added purpose of helping to highlight the potential role of professional bodies in promoting and supporting data management skills development amongst professionals in their fields.

DaMSSI also worked alongside the RIN Information-Handling Working Group on a number of initiatives, including contributions to a taxonomy for an information-handling 'lens' for the RDF, which will be published in late 2011.

5. RECOMMENDATIONS AND FUTURE DEVELOPMENT

DaMSSI has drawn together the following list of recommendations for future providers of data management training, based on its findings:

Work closely with disciplinary experts to ensure that terminology used within courses is accurate and clear, including agreeing a basic definition of core concepts such as what 'data' can be within the discipline;

Keep overviews and central descriptions of topic areas basic and generic, introducing the topic at a digestible level and allowing for easier integration into existing larger research methods courses;

Interlace generic with discipline-specific examples, references and case studies wherever possible, highlighting relevance, engaging audience and putting basic points into context;

Translate jargon for the audience, explaining principles and issues in language researchers/students can understand;

Offer access to customised DMP guidance for the discipline so students can produce plans specifically relevant to them;

Have trainers with extensive knowledge of the discipline, who can provide the context and interlaced examples that engage students and make the topic seem relevant to them;

Offer training in the basic principles of data management at an early stage in postgraduate studies, allowing students to begin their project using best practice;

Be concise, with basic modules short enough to maintain interest and be integrated into larger research skills courses;

Deliver face-to-face training, as attendees find the opportunity to exchange experiences and thoughts with others invaluable. However students also want access to online training materials for ongoing reference and for those unable to attend courses in person;

Stress the potential benefits associated with good data management practice, such as helping researchers to secure funding and meeting legal requirements;

Work with professional bodies and funders to endorse and promote good data management practice, helping students and researchers to have support and potential reward for their efforts from leaders and funders within their discipline.

The work begun by DaMSSI is now being taken forward by the RIN and DCC through the analysis of longer term data management skills development for specific disciplines, and of current UK LIS courses against the skills identified by DaMSSI and the RDMTrain projects. There have been expressions of interest in extending the suite of career profiles by the DCC, RIN, professional bodies and some international partners. The EU-funded DigCurV project may also incorporate the findings of DaMSSI into their design and development of a digital curation training curriculum.

6. ACKNOWLEDGMENTS

Our sincere thanks to Dr Simon Hodson, JISCMRD programme manager; Joy Davidson at HATII/DCC; Stéphane Goldstein at RIN and all the JISC RDMTrain projects for their support and input.

7. REFERENCES

[1] National Science Foundation. 2005. *Long-lived digital data collections: enabling research and education in the 21st century*. (Sep. 2005). http://www.nsf.gov/pubs/2005/nsb0540/.

[2] JISC Data Skills and Capacity session at the JISC Innovation Forum 2008 (Keel, 14-15 July 2008). http://jif08.jiscinvolve.org/wp/theme-2-the-challenges-of-research-data/sub-page-2/.

[3] Ward, C., Freiman, L., Jones, S., Molloy, L. and Snow, K. 2010. *Incremental Scoping Study and Implementation Plan*. (July 2010). http://www.lib.cam.ac.uk/preservation/incremental/documents/Incremental_Scoping_Report_170910.pdf.

[4] The Research Data Management Training Materials strand is described at: http://www.jisc.ac.uk/whatwedo/programmes/mrd/rdmtrain.aspx.

[5] SCONUL Working Group on Information Literacy. 2011. *The Seven Pillars of Information Literacy Model: Research Lens for Higher Education.* (April 2011). http://www.sconul.ac.uk/groups/information_literacy/sp/publications/researchlens.pdf.

[6] Vitae. 2009. *The Researcher Development Framework.* (2009). http://www.vitae.ac.uk/CMS/files/upload/Vitae-Researcher-Development-Framework.pdf.

[7] The mapping of of RDMTrain projects against the RDF can be found at http://www.rin.ac.uk/data-management-skills.

Capitalizing on the State-of-the-Art in Preserving Complex Visual Digital Objects: The POCOS Project

Leo Konstantelos
HATII at the University of Glasgow
11 University Gardens
Glasgow G12 8QH, UK
+44 (0) 141 330 7133
Leo.Konstantelos@glasgow.ac.uk

David Anderson
Janet Delve
Milena Dobreva
Clive Billenness
Future Proof Computing Group
School of Creative Technologies
University of Portsmouth
004423 9284 5491
{david.anderson; janet.delve; milena.dobreva}@port.ac.uk
Clive.Billenness@bl.uk

Richard Beacham
Drew Baker
King's Visualisation Lab (KVL)
Centre for Computing in the Humanities (CCH),
King's College London, 26-29 Drury Lane, London WC2B 5RL, UK
{richard.beacham, drew.baker}@kcl.ac.uk

Vincent Joguin
Sonia Séfi
Joguin sas
30 chemin du Vieux Chêne
38240 Meylan, France
{Vincent, sonia.sefi}@joguin.com

ABSTRACT

Complex visual digital objects and environments present the digital preservation community with distinct challenges. Complex visual objects predominantly feature interactivity properties, time-based components and intricate interdependencies which incorporate composite, heterogeneous, and often bespoke technologies. Work recently undertaken in major EC-funded projects has highlighted that continuing progress in preserving complex digital materials is achievable through engagement with relevant communities and amalgamation of research results and emerging good practices. The JISC-funded project Preservation of Complex Objects Symposia (POCOS) addresses these issues by creating a template for leading researchers and practitioners to present their findings and set out the future directives in this field. To this end, POCOS will deliver a series of three symposia focusing on three respective areas: Simulations and Visualizations, Software-based Art and Gaming Environments and Virtual Worlds. POCOS aims to promote broader appreciation of the state-of-the art in preserving complex objects, provide input to collections management and create fertile ground for future collaborations between academia and industry.

Categories and Subject Descriptors

J.5 [**Computer Applications**]: Arts and Humanities – *Fine Arts*
K.4 [**Computers and Education**]: Computers and Society – *Miscellaneous*
K.8 [**Personal Computing**]: General - *Games*

General Terms

Management, Documentation, Human Factors, Standardization, Theory, Legal Aspects

Keywords

complex visual objects, digital preservation, software art, simulations and visualizations, gaming environments.

1. INTRODUCTION

As the use of digital resources has started to include more complex structures and environments, digital curation and preservation professionals are confronted by the intellectual and logistical challenges that this shift implies. Complex visual digital objects have moved beyond the experimental sphere: they are becoming increasingly central to learning and research environments [e.g. 1, 2, 3], and their role as records of modern culture is equally recognized by heritage and memory institutions [e.g. 4, 5].

The challenges posed to digital preservation can be witnessed in a number of fronts. Technically, complex visual digital objects feature interactivity properties, time-based components and intricate interdependencies which incorporate composite, heterogeneous, and often custom-built technologies. Digital preservation has thus far predominantly focused on migration-based approaches for core digital formats – for instance, research data and textual representations – whose applicability is arguable when dealing with the complexity of materials such as video games and three dimensional virtual worlds. Conceptually, complex, born-digital visual artifacts deviate from traditional (analogue) media and are difficult to describe and document in a formalized manner. As previous studies have shown [e.g. 6, 7] the results from applying existing archival theory and metadata standards to complex objects can be highly variable. Attempting to include born-digital visual objects (or at least references to them) in structured repositories and digital libraries can introduce further complications [8]. Historically, complex digital materials are a relatively young medium within artistic, educational and

research contexts and therefore the deriving artifacts "do not have that history of production and scholarship, nor is there time to 'hope for the best' in terms of preservation" [9]. Under this prism, it is imperative to provide a template for synergies between researchers and practitioners in otherwise disparate fields, thus generating knowledge economies and ultimately contributing to a collective memory in the area of complex digital objects so that "it can benefit from remembering its past more systematically." [10]

These are the issues that the Preservation of Complex Objects Symposia (POCOS)[1] project addresses. With funding from the Joint Information Systems Committee (JISC)[2], POCOS will deliver a series of three symposia across the United Kingdom with the aim to bring together the leading researchers and practitioners in the field and invite them to present their findings, identify key unsolved problems, and map out the future research agenda for the preservation of complex visual materials and environments. The fundamental task facing the POCOS symposia is to present material of great technological and organizational complexity in a lucid, cogent, relevant and approachable manner so as to engage HEIs' researchers and practitioners in a wide variety of disciplines, as well as reaching those further afield in, for example, commerce, industry, cinema and government. The ultimate goal of POCOS is to promote a broader appreciation of the state-of-the art in preserving complex objects, provide input to collections management and create fertile ground for future collaborations between academia and industry.

2. POCOS SYMPOSIA

Throughout 2011 POCOS will hold symposia in London, Glasgow and Cardiff. Although each symposium is organised by a specific project partner, the interrelations among the three domains of focus and their respective digital preservation issues further promote the definition of good practice guidelines in a collaborative manner. The three non-orthogonal areas of the symposia highlight the potential of synergies in preserving complex objects, allowing at the same time for deliberations on digital preservation theory and practice that is specific to each area.

2.1 Simulations & Visualizations

Proprietary data formats, proprietary methods and processes, and inaccurate data structures [11] are just a few of the problems pertaining 3D modelling. Successful preservation of 3D models depends on a number of parameters, including identification of file formats, specification of technical characteristics and standardisation of metadata models. Furthermore, accurate interpretation of 3D models depends on the persistence of the software used to create, render and display the deriving products. Through presentation of real-life case studies, focused discussion and networking activities the POCOS symposium on Simulations and Virtualisations deals with such issues as intellectual transparency in 3D cultural heritage material, the role of virtual museums and preservation of mixed reality representations of heritage sites. The symposium is organised by the King's Visualisation Lab[3] based at the Centre for Computing in the Humanities, King's College London.

2.2 Software Art

The integration and manipulation of technology as a form of artistic expression has been an active and growing genre for more than fifty years. Software-based artworks have been commissioned and displayed in major museums across the globe, therefore emphasising on the need to curate, manage and preserve such material. Preservation of software-based art presents challenges in many fronts, including complex interdependencies between objects; time-based and interactive properties; and diversity in the technologies and practices used for development [12]. Although some guidelines exist for preserving and curating software-based artworks [e.g. 13], there currently exist no agreed upon methods and techniques that can broadly constitute 'good practice'. With contributions from artists, software engineers, museum and gallery curators, as well as representatives from academia and research, the POCOS symposium on software art addresses such topics as the implications and advances in preserving software-based art, issues of ephemerality, significant properties for software-based art, connections with software preservation in general, and software-based art as *performance*. The symposium is organised by the Humanities Advanced Technology & Information Institute[4] at the University of Glasgow.

2.3 Gaming Environments

Video games have been a prominent feature of popular culture with a history than spans more than five decades. Similarly, prototypical implementations of virtual worlds appeared as early as 1974 and have nowadays exploded into such phenomena as Massively multiplayer online role-playing games (MMORPGs) and Second Life[5]. Despite their role in shaping and expressing socio-cultural identity, preservation of gaming environments and virtual worlds did not gain serious attention until recently. The POCOS symposium on Gaming Environments and Virtual Worlds brings together outstanding game developers, virtual worlds producers, academics, heritage institutions and members of the experimental gaming community in an effort to synchronise their actions towards preserving their assets and reach an understanding in terms of best practices, legal implications and future directives. Key topics include digital games preservation and exhibitions, digital games/virtual worlds history and documentation, computer demos preservation and their alignment with software preservation in general. The symposium is organized by Joguin sas, France[6].

3. PROJECT OUTPUTS

The POCOS project will release selected content from the symposia as video footage and audio on popular media sharing platforms, so as to communicate the deliberations to the broader international community and generate material of lasting value. After conclusion of each symposium, a peer-reviewed publication

[1] http://www.pocos.port.ac.uk/

[2] http://www.jisc.ac.uk/

[3] http://www.kvl.cch.kcl.ac.uk/

[4] http://www.gla.ac.uk/departments/hatii/

[5] http://secondlife.com/

[6] http://www.joguin.com/

will be prepared with key texts that encapsulate the content of the POCOS events and provide concrete recommendations and pointer for future directions. The publications will be disseminated in a variety of electronic means and retained in an open-access institutional repository.

4. OUTCOMES AND FUTURE WORK

POCOS envisages delivering results within the three domains of focus and also providing input to the greater digital preservation field. As it has been exhibited throughout this paper, the project aims to stimulate a broader appreciation of the state-of-the art research in the area of preservation of complex digital objects, working at the same time towards a consensus on potential future avenues of research and practice. The corpus of published material is meant to provide input to the strategic planning of holders of collections of complex materials and environments, helping institutions to synchronise their practices between core and complex objects. By exploiting the possibilities that interrelations between the domains of focus offer, POCOS seeks the creation of new research networks to pursue such research and harness its outputs in a coordinated and more cost-effective manner. As opposed to previous efforts that have focused on technical and conceptual issues pertaining to complex visual digital objects, POCOS suggests that, in order to continue to make progress, it is important to engage and energize the wider DP community. This paper has summarized the key topics from which the POCOS project was forged, and has demonstrated the importance and timeliness of this work.

5. ACKNOWLEDGMENTS

Preservation Of Complex Objects Symposia (POCOS) is coordinated by the University of Portsmouth and managed by the British Library. The project is supported with funding by the Joint Information Systems Committee (JISC).

6. REFERENCES

[1] Baranowski, T., Buday, R., Thompson, D. I., & Baranowski, J. (2008). Playing for Real: Video Games and Stories for Health-Related Behavior Change. American Journal of Preventive Medicine, 34(1), 74-82.

[2] Beacham, R. (2008). “Oh, to make boards to speak! There is a task!": Towards a Poetics of Paradata. In M. Greengrass and L. Hughes (Eds). The Virtual Representations of the Past (pp. 171-178). Aldershot, Hants, England ; Burlington, VT : Ashgate.

[3] Konstantelos, L. (2009). Digital art in digital libraries: a study of user-oriented information retrieval. PhD thesis, University of Glasgow.

[4] Paul, C. (2008). New media in the white cube and beyond: Curatorial models for digital art. Berkeley: University of California Press.

[5] Manovich, L. (2003). Making art of databases. Rotterdam: V2 Pub./NAi Pub.

[6] Rinehart, R. (2004). A System of Formal Notation for Scoring Works of Digital and Variable Media Art. Paper presented at the Electronic Media Group, Annual Meeting of the American Institute for Conservation of Historic and Artistic Works, Portland, Oregon.

[7] Depocas, A., Ippolito, J., & Jones, C. (2003). Permanence through change: The variable media approach. New York: Guggenheim Museum Publications.

[8] Konstantelos, L. (2007). Putting the content, user and quality concepts in the digital library universe into practice: a scenario based on a user-oriented study for digital art material. In: Castelli, D. and Ioannidis, Y. (eds.) Proceedings of the Second Workshop on Foundations of Digital Libraries. Information Society Technology Press, pp. 22-30.

[9] Dalbello, M., Marty, P., Paling, S., Simon, S., Walsh, J., Winget, M., et al. (2008). Mapping work in the arts and humanities: A participatory panel discussion. Proceedings of the American Society for Information Science and Technology, 45(1), 1-3.

[10] Manovich, L. (2003). Don't Call It Art: Ars Electronica 2003. Retrieved March 16, 2011, from http://www.manovich.com/DOCS/ars_03.doc

[11] Vilbrandt, C., Pasko, A., Pasko, G., Goodwin, J.R., & Goodwin, J.M. (2001). Digital Preservation of Cultural Heritage through Constructive Modeling. Proceedings of the International Cultural Heritage Informatics Meeting (ichim01), Milano, Italy, September 3-7, 2001.

[12] Cox, G. (2010). Antithesis: the Dialectics of Software Art. Aarhus, DK: Digital Aesthetics Research Center.

[13] Boudrias, E., Mingarelli, A., Driss, O., and Gangier, R. (2009). A Preservation Guide for Technology-Based Artworks. Available from the Documentation and Conservation of the Media Arts Heritage (DOCAM): http://www.docam.ca/en/conservation-guide.html

Building Digital Preservation Practices, Tools and Services on Quicksand

Bram van der Werf
Open Planets Foundation
The British Library, Boston Spa
Wheterby, West Yorkshire
+31 6 424 06 775
bram@openplanetsfoundation.org

QUICKSAND: THE FUNDING MODEL

The existence of today's digital preservation tools and services within the cultural heritage sector proves there is a need for them and that financial resources can be made available to develop functionalities that address specific digital preservation issues. However, what is really lacking is a solid business case for maintaining these tools over time.

It is best to look at the IT industry to see how the maintenance of products and services works. The life-cycle of a major operating system, for example, is four to six years, with two to three major releases during this period. The releases are part of the maintenance effort and consist of bug fixes and implementation of cost saving and/or innovative improvements. The release and maintenance strategy of operating system vendors has in its turn a strong impact on the cost model for the services and applications that run on top of these systems. These services and applications need to follow the pace of maintenance of the underlying software layer and at the same time they also follow their own bug fixing/improvement cycles. These accumulated maintenance cycles add up exponentially. In general it can be stated that approximately 20% (or less) of the total cost of almost any application or service goes into its development and over 80% (or more) of the total cost is required for its maintenance.

Project-funded software development does not consider maintenance costs. If the aim of a project is solely to produce a software product with specific functionality, and it does not consider the responsibility to maintain this functionality during the life-cycle of the product, then it is very unlikely that this product will be sustainable. Not only maintenance issues are at stacks, even worse, project planning and financial constraints inhibit proper software development. Indeed, in order to deliver software which meets the criteria defined in the project plan, project owners and developers will often be tempted to use approaches that do not really take deployability, long-term maintainability and supportability into consideration. This applies to both the technology and the content back-end which feeds the system with up-to-date information. The content back-end is a crucial source of information (usually a database filled by human intelligence and maintained by human effort) for the application. Typical examples of such applications that are dependent on content back-end are DNS-services and persistent identifier resolver mechanisms. In addition, software technologies are often based on niche and weakly supported technologies, whereas the information backend lacks a competent community to feed it with up to date information.

What we see in the cultural heritage sector happening in the past two decades, is a flourishing R&D activity, based on generous project funding and hardly any serious commitment for deployment, maintenance and sustainability. The sector operates a digital preservation and long-term access business process with tools, prototypes and services that lack any appropriate long-term business planning. The global community of heritage institutions seems unable to secure appropriate structural funding for long-term commitment to digital preservation. Most efforts heavily depend on projects funding and are because of that under the heavy scrutiny of political and financial climates.

THE REAL PROBLEM

Many years of R&D effort have been invested in digital preservation, and more importantly, in developing tools and services to aid long-term access to digital material. In practice, however, those responsible for preservation and long-term access are confronted with the urgency to take ad hoc actions that respond to 1) content production and distribution trends and 2) the needs of users accessing the content. There is still a large gap between R&D and practice. This has been the same for ages, regardless of digital or analogue content. It has always been important to preserve media, whether clay tablets, papyrus, paper or more recently bits and bytes on digital carriers. Understanding context and content (e.g the rosetta stone),in other words keeping the contextual and content metadata, is equally important. The challenge of preserving digital information requires the same tiered conceptual thinking. Bit preservation and media obsolescence remains a risk and though this in itself is not a trivial challenge, understanding context and content of digital objects is a far greater challenge in terms of its risk and mitigation level. Perhaps it helps to articulate long-term access as the real challenge, rather than bit preservation. Bit preservation and longevity of bit quality and integrity requires resources and often implies buying power, whereas long-term access requires a range of reliable and stable information sources and availability of competent and dedicated human resources supported by analysis and decision-making tools, allowing organizations to test, verify and decide on actions for managing accessibility to digital objects over time.

iPRES2011, Nov. 1–4, 2011, Singapore.

DIGITAL AS PANACEA

The majority of society assumes that preserving digital information is similar to preserving paper copies or should at least be cheaper and easier – certainly not more costly and complex. The use of digital media has been marketed to users as utilities offering them higher volumes of information with easier and faster access, but marketing has almost never focused on the associated risks. Where analogue media can be accessed and preserved autonomously, digital objects have a strong dependency on rendering software. The wider community lacks an understanding of technology and does not realize that rendering digital information also requires the right software. Society is also used to the longevity of analogue media (paper, vinyl) and therefore lacks a sense of urgency to take timely action to preserve digital media. Access to information held on CD, DVD, USB sticks etc. is only guaranteed during the commercial life-span of such carriers, which is deceptively short.

For software and hardware vendors, their primary objective is to run a healthy and profitable business. As long as their products generate revenue, they have neither business objective nor commercial interest to guarantee long-term access to digital objects produced for non-supported software versions - unless it generates profit. Where the academic and cultural heritage sectors think in terms of centuries for preservation, technology vendors and format producers often don't even think in terms of decades.

Beyond the cultural heritage and academic sector, awareness of the importance of long-term access to digital objects and the complexity of it, is low to non-existent. Over the past few decades, technology vendors have managed to convince society of "digital" being the panacea for our insatiable need for lots of open, easily accessible and user-friendly information.

Societal naivety and commercial interests are not helpful to convey the necessary sense of urgency and to enforce structural financial commitment for long-term access solutions

NO FREE RIDE

But even in the cultural heritage and academic sector, explaining the consequences of long-term access and taking adequate measures accordingly proves difficult. By contrast, explaining the importance of immediate access and of building information portals is an easy win when needing to convince decision-makers and funders. Open, reliable and user-friendly experiences as we perceive today can be well articulated and marketed as products. As immediate results, they are a good fit for commercial exploitation or fancy demos.

Maintenance services that are necessary to manage these experiences over time are however much harder to market and reactions to such services are more similar to how society reacts to maintenance, insurance and other less tangible, not directly product related activities. When we think about flexible transportation for example, we experience the idea of buying a car as a very positive thing. The price of the car and even the accessories are perceived as fully acceptable, but the pain comes with insurance, maintenance, repairs, taxes, tires etc. It is the big bad world that wants to spoil our joy and only when things go wrong do we appreciate the value of some of these services.

And often during tendering and other types of project or program negotiations, it is on maintenance costs that cost cutting takes place. On an operational level long-term access is a technology challenge with a typical products vs service equation in terms of financial consequences. A 20/80 % cost consideration, 20% for developing a software solution, versus 80% for keeping this solution well maintained and up to date over a period of 3-4 years.

Most of today's information services are being developed with project funding or sponsored by single organizations. Commercial digital preservation solutions actually use information from these services due to the non-existence of business models for sustainable information services with an SLA option and the ability to provide reliable information required for assuring long-term access. This is actually where the term "no free ride" gets put into practice as well. In both the private and public sector, no sustainable service survives without a decent business plan, so therefore tools and services that support digital preservation and long-term access can and will never be a "free ride"

WHAT NEXT?

R&D, innovation and organizational initiatives over the last 12-15 years have produced several long-term access products and services. Many of them have a functional potential to become of significant value and to be mechanical to long-term access. Most of them originated from project funding (Mellon, EC, JISC, NDIIP) or incidental program or project funding coming from individual organizations. Looking back, while appreciating the initial functionality and quality of many of these services, most of them lack long-term sustainability when it comes to maintaining the initial information quality and when it comes to its capacity to adapt to change. Often this is caused by the fact that it takes a relatively big economical effort to monitor, edit and maintain all the potential information resources.

Information coming from these services or architectures is almost trivial compared to the fact that in reality it takes a network of competent, dedicated people, their input and above all a business model with a structural funding, to maintain the quality of the content of such a service over time.

An Open Source approach where a community is actively supported and financed by stakeholders willing to dedicate competent resources is a feasible and realistic option, as business model. But this type of community also needs moderation, leadership and long-term funding to be able to steward, sustain, maintain and manage the solution in the interest of the same stakeholders. While being a great option in theory, practice is more complicated. Sense of community is based on accepting commonalities and the biggest players in ALM sector still tend to foster differences. Another complicating factor is mistrust towards technology service providers while there is a high need for some of their core competences such as engineering skills.

Other business models can be services by subscription, API's with license keys connected to payments, Software as a Service (SaaS) with contracts based on volume or size of organization, or a community model financed by stakeholders and users similar to the DOI business model for persistent identifiers.

Subscription, license, SaaS and API type of business models all assume that core knowledge will reside outside organizations, it is for this very reason that a strong community endorsing and sharing Open Source solutions will bring most value and long term sustainability of solutions to Libraries and Archives. But this has to be treated as a funded business model and not a "free ride".

AUTHOR INDEX

Alter, George 215
Altman, Micah 256
Anderson, David 44, 282
Antunes, Gonçalo 1, 81
Baker, Drew 282
Barateiro, José 1, 70
Beacham, Richard 282
Béchard, Lorène 11
Becker, Christoph 1, 52
Belhajjame, Khalid 228
Bergmeyer, Winfried 44
Billenness, Clive 282
Bøgvad, Ulla Kejser 107
Borglund, Erik A.M. 205
Bos, Marguérite 120
Brown, Geoffrey 181
Canteiro, Sara 70
Carr, James 267
Castro, Rui 252
Cho, Won-Ik 271
Christian, Thu-Mai 256
Ciuffreda, Antonio 44, 258
Cochrane, Euan 148
Conway, Esther 276
Crabtree, Jonathan 242, 256
Dappert, Angela 33
David, Gabriel 265
De Roure, David 228
Delve, Janet 44, 258, 282
Dobreva, Milena 282
Dokkedal, Barbara 24
Donaldson, Devan Ray 20
Double, Jodie 232
Draws, Daniel 130
Edelstein, Orit 194
Esteva, Maria 93
Euteneuer, Sven 130
Evans, Mark 267
Factor, Michael 194
Faria, Luís 252
Ferreira, Miguel 252
Foscarini, Fiorella 260
Freitas, Ricardo André Pereira 140
Gerber, Urs 120
Goble, Carole 228
Goebert, Samuel 254
Gómez-Pérez, José Manuel 228
Graf, Roman 190
Grindley, Neil 116
Guttenbrunner, Mark 171, 259
Hamm, Markus 52
Hettne, Kristina 228
Hoeven, Jeffrey van der 167
Huber-Mörk, Reinhold 190
Innocenti, Perla 269
Jackson, Andrew N. 33, 89, 232
Joguin, Vincent 44, 282
Johnston, Leslie 210
Jordan, Christopher 93
Kiers, Bart 167
Kim, Young-Joo 271
Kim, Yunhyong 262
Kimura, Akiko 33
King, Ross 194
Kirstein, Michael 120
Klyne, Graham 228
Konstantelos, Leo 44, 258, 269, 282
Lambert, Simon 276
Lange, Andreas 44
Lee, Kwan-Yong 271
Lohman, Bram 167
Lopes, João Correia 224
Lyle, Jared 215
Mason, Paul 120
Massol, Marion 11
Matthews, Brian 276
McGovern, Nancy 256
McGuinness, Rebecca 232
Michel, David 167
Middleton, Bo 232
Missier, Paolo 228
Molloy, Laura 279
Naumann, Kai 120
Neuroth, Heike 274
Nielsen, Anders Bo 107
Nielsen, Anders Bo 24
Oliver, Gillian 260
Osswald, Achim 274
Oury, Clément 237
Palma, Raúl 228
Peyrard, Sébastien 237
Pient , Amy 215
Pierson, Jason 267
Pina, Helder 81
Pinchbeck, Dan 44
Rahman, Arif Ur 265
Ramalho, José Carlos 140
Rauber, Andreas 62, 97, 171, 250
Rechert, Klaus 158, 246, 248
Ribeiro, Cristina 224, 265
Risse, Thomas 194
Rönsdorf, Carsten 120
Roos, Marco 228
Ross, Seamus 262, 269
Rouchon, Olivier 11
Ruiz, José Enrique 228
Salant, Eliot 194
Samuelsson, Göran 120
Sarti, Alain 254
Schmelzer, Sebastian 246
Séfi, Sonia 282
Shaon, Arif 120, 276
Sharpe, Robert 267
Silva, João Rocha da 224
Simon, Daniel 130
Simon, Frank 130
Snow, Kellie 279
Strathmann, Stefan 274
Strodl, Stephan 97
Suchodoletz, Dirk von 148, 158, 246, 248
Taylor, Philip 194
Thirifays, Alex 24, 107
Urban, Tomislav 93
Valizada, Isgandar 158, 248
Vieira, Ricardo 1
Walling, David 93
Weihs, Christian 62
Werf, Bram van der 285
Wheatley, Paul 232
Woolf, Andrew 120