iPRES 2016

iPRES 2016

13th International Conference on Digital Preservation //

Proceedings

Bern // October 3 – 6, 2016

www.ipres2016.ch

Proceedings of the 13th International Conference on Digital Preservation

Swiss National Library
Hallwylstrasse 15
CH-3003 Bern

www.nb.admin.ch

Hosted by:

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Eidgenössisches Departement des Innern EDI
Département fédéral de l'intérieur DFI
Dipartimento federale dell'interno DFI
Federal Department of Home Affairs FDHA
Schweizerische Nationalbibliothek NB
Bibliothèque nationale suisse BN
Biblioteca nazionale svizzera BN
Swiss National Library NL

Sponsored by:

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Eidgenössisches Departement des Innern EDI
Département fédéral de l'intérieur DFI
Dipartimento federale dell'interno DFI
Federal Department of Home Affairs FDHA
Schweizerische Nationalbibliothek NB
Bibliothèque nationale suisse BN
Biblioteca nazionale svizzera BN
Swiss National Library NL

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Schweizerisches Bundesarchiv BAR
Archives fédérales suisses AFS
Archivio federale svizzero AFS
Swiss Federal Archives SFA

CERN

Google

ETH BIBLIOTHEK

ExLibris
a ProQuest Company

swissuniversities

iPRES 2016
Bern // October 3 – 6, 2016

Proceedings of the 13th International Conference on Digital Preservation

TABLE OF CONTENTS //

Preface........ 9
Conference Organisation........ 12

Short Papers // 14

A Decade of Preservation: System Migrations in Chronopolis
Sibyl Schaefer, Mike Smorul, David Minor, Mike Ritter........ 15

Digital Preservation through Digital Sustainability
Matthias Stuermer, Gabriel Abu-Tayeh........ 18

Building a Future for our Digital Memory: A National Approach to Digital Preservation in The Netherlands
Marcel Ras........ 23

Status of CELLAR: Update from an IETF Working Group for Matroska and FFV1
Ashley Blewer, Dave Rice........ 27

PREMIS 3.0 Ontology: Improving Semantic Interoperability of Preservation Metadata
Angela Di Iorio, Bertrand Caron........ 32

Exhibiting Digital Art via Emulation - Boot-to-Emulator with the EMiL Kiosk System
Dragan Espenschied, Oleg Stobbe, Thomas Liebetraut, Klaus Rechert........ 37

Project „The Digital City Revives": A Case Study of Web Archaeology
Tjarda de Haan........ 41

Ellipse – Long-term and Permanent Protection and Accessibility of Geodata
Tobias Mérinat, Chiara Marciani, Krystyna Ohnesorge, Alain Mast........ 46

Preserving Research Data: Linking Repositories and Archivematica
Jenny Mitcham, Matthew Addis, Chris Awre, Julie Allinson, Richard Green, Simon Wilson........ 50

Implementing Automatic Digital Preservation for a Mass Digitization Workflow
Henrike Berthold, Andreas Romeyke, Jörg Sachse, Stefan Fritzsche, Sabine Krug........ 54

Applied Interoperability in Digital Preservation: solutions from the E-ARK Project
Kuldar Aas, Janet Delve, Andrew Wilson........ 57

Copyright in Cultural Heritage Institutions: Snapshot of the Situation in Switzerland from a librarian perspective
Andrea Ruth Schreiber........ 62

The Ties that Bind - On the Impact of Losing a Consortium Member in a Cooperatively Operated Digital Preservation System
Michelle Lindlar........ 66

A Method for Acquisition and Preservation of Email
Claus Jensen, Christen Hedegaard........ 72

Processing Capstone Email Using Predictive Coding
Brent West, Joanne Kaczmarek........ 77

Using RMap to Describe Distributed Works as Linked Data Graphs: Outcomes and Preservation Implications
Karen Hanson, Sheila Morrissey, Aaron Birkland, Tim Dilauro, Mark Donoghue........ 80

Long Papers // 84

An OAIS-oriented System for Fast Package Creation, Search, and Access
Sven Schlarb, Rainer Schmidt, Roman Karl, Mihai Bartha, Jan Rörden, Janet Delve, Kuldar Aas 85

Securing Trustworthy Digital Repositories
Devan Ray Donaldson, Raquel Hill, Heidi Dowding, Christian Keitel........ 95

Will Today's Data Be Here Tomorrow? Measuring The Stewardship Gap
Jeremy York, Myron Gutmann, Francine Berman........ 102

What Makes a Digital Steward: A Competency Profile Based On The National Digital Stewardship Residencies
Karl-Rainer Blumenthal, Peggy Griesinger, Julia Kim, Shira Peltzman, Vicky Steeves........ 112

Identifying Barriers To File Rendering In Bit-level Preservation Repositories: A Preliminary Approach
Kyle Rimkus, Scott Witmer........ 121

Practical Analysis of TIFF File Size Reductions Achievable Through Compression
Peter May, Kevin Davies........ 129

Towards a Risk Model for Emulation-based Preservation Strategies: A Case Study from the Software-based Art Domain
Klaus Rechert, Patricia Falcao, Tom Ensom........ 139

A Case Study on Emulation-based Preservation in the Museum: Flusser Hypertext
Frank Padberg, Philipp Tögel, Daniel Irrgang, Martin Häberle........ 149

Precise Data Identification Services for Long Tail Research Data
Stefan Proell, Kristof Meixner, Andreas Rauber........ 159

CERN Services for Long Term Data Preservation
Jamie Shiers, Frank Berghaus, German Cancio Melia, Suenje Dallmeier Tiessen, Gerado Ganis, Tibor Simko, Jakob Blomer........ 168

Designing Scalable Cyberinfrastructure for Metadata Extraction in Billion-Record Archives
Gregory Jansen, Richard Marciano, Smruti Padhy, Kenton McHenry........ 177

Navigating through 200 Years of Historical Newspapers
Yannick Rochat, Maud Ehrmann, Vincent Buntinx, Cyril Bornet, Frédéric Kaplan........ 186

Towards a Systematic Information Governance Maturity Assessment
Diogo Proença, Ricardo Vieira, Jose Borbinha........ 196

Conceptualising Optimal Digital Preservation and Effort
Sean Mosely, Jessica Moran, Peter McKinney, Jay Gattuso........ 206

Mapping Significance of Video Games in OAIS
Rhiannon Bettivia........ 212

Database Preservation Toolkit: A relational database conversion and normalization tool
Bruno Ferreira, Luis Faria, José Carlos Ramalho, Miguel Ferreira........ 222

Exploring Friedrich Kittler's Digital Legacy on Different Levels: Tools to Equip the Future Archivist
Heinz Werner Kramski, Jürgen Enge ... 229

Persistent Web References – Best Practices and New Suggestions
Eld Zierau, Caroline Nyvang, Thomas Kromann ... 237

Preserving Websites of Research & Development Projects
Daniel Bicho, Daniel Gomes ... 247

Posters // 255

01 A Demonstration of BitCurator Access Webtools and Disk Image Redaction Tools
Christopher Lee, Kam Woods ... 256

02 Assigning Rights Statements to Legacy Digital Collections
Laura Capell, Elliot Williams ... 258

03 Autonomous Preservation Tools in Minimal Effort Ingest
Asger Askov Blekinge, Thorbjørn Ravn Andersen, Bolette Ammitzbøll Jurik ... 260

04 Born Digital 2016: Generating public interest in digital preservation
Sarah Slade ... 262

05 Building Archival Storage Component Using New Open Standard Infrastructure
Dariusz Paradowski ... 264

06 Caring For Live Art That Eludes Digital Preservation
Louise Curham ... 266

07 Consortial Certification Processes – The Goportis Digital Archive. A Case Study
Yvonne Friese, Thomas Gerdes, Franziska Schwab, Thomas Bähr ... 268

08 Digital Preservation with the Islandora Framework at Qatar National Library
Armin Straube, Arif Shaon, Mohammed Abo Ouda ... 270

09 Establishing a generic Research Data Repository: The RADAR Service
Angelina Kraft, Matthias Razum, Jan Potthoff, Ena Brophy ... 272

10 Establishing Digital Preservation At the University of Melbourne
Jaye Weatherburn ... 274

11 Exit Strategies and Techniques for Cloud-based Preservation Services
Matthew Addis ... 276

12 Geospatial Data: Conservation and Archiving Planning
Helen Gollin, Urs Gerber, Martin Schlatter ... 278

13 HydraDAM2: Building Out Preservation at Scale in Hydra
Heidi Dowding, Michael Muraszko ... 280

14 Project EMiL – Emulation of Multimedia Objects
Tobias Steinke, Frank Padberg, Astrid Schoger, Klaus Rechert ... 282

15 Should We Keep Everything Forever?: Determining Long-Term Value of Research Data
Bethany Anderson, Susan Braxton, Elise Dunham, Heidi Imker, Kyle Rimkus ... 284

16 The Status of Compliance with OAIS Reference Model in the National Diet Library Digital Collections
Tsukasa Kimezawa, Shuji Kamitsuna ... 286

17 To Act or Not to Act – Handling File Format Identification Issues in Practice
Matthias Töwe, Franziska Geisser, Roland Suri ... 288

18 Web Archiving Environmental Scan
Gail Truman, Andrea Goethals ... 290

Panels // 292

OAIS for all of us
Marcel Ras, William Kilbride, Sabine Schrimpf, Barbara Sierman ... 293

Software Sustainability and Preservation: Implications for the Long-Term Access to Digital Heritage
Euan Cochrane, Jessica Meyerson, Natasa Milic-Frayling, David Rosenthal, Zach Vowell ... 294

Challenges and benefits of a collaboration of the Collaborators: How can we learn and benefit from experiences and set next steps in international collaboration
Marcel Ras, William Kilbride, Sabine Schrimpf, Joachim Jung ... 296

Workshops & Tutorials // 298

Workshop 01 What is Preservation Storage?
Andrea Goethals, Steve Knight, Jane Mandelbaum, Nancy McGovern, Gail Truman, Kate Zwaard, Sibyl Schaefer ... 299

Workshop 02 Personal Digital Archiving: How Can the Cultural Heritage Community Help Individuals Curate Their Own Materials?
Maurizio Lunghi, Helen Tibbo, Christopher Lee, Natascha Schumann, Achim Osswald, Martin Iordanidis ... 300

Workshop 03 Quality Standards for Preserving Digital Cultural Heritage
Claudio Prandoni, Antonella Fresa, Börje Justrell, Peter Pharow, Magnus Geber, Erwin Verbruggen ... 302

Workshop 04 OSS4Pres 2.0: Building Bridges and Filling Gaps
Sam Meister, Carl Wilson, Shira Peltzman, Heidi Dowding ... 304

Workshop 05 Innovative Practices in Digital Preservation, Data Life Cycle Management (DLCM): A Swiss Initiative
Pierre-Yves Burgi, Eliane Blumer, Aude Dieudé, Ana Sesartic ... 306

Workshop 06 Symmetrical Web Archiving with Webrecorder
Dragan Espenschied, Ilya Kreymer ... 307

Tutorial 07 Persistent Identifiers for Digital Cultural Heritage
Jonathan Clark, Maurizio Lunghi, Remco Van Veenendaal, Marcel Ras, Juha Hakala ... 309

Workshop 08 Relational Database Preservation Standards and Tools
Luis Faria, Marcel Büchler, Kuldar Aas ... 311

Tutorial 09 Understanding and Implementing PREMIS
Peter McKinney, Eld Zierau, Evelyn McLellan, Angela Dappert ... 313

Tutorial 10: Introduction to Fedora 4
David Wilcox, Andrew Woods.......... 315

Workshop 11 How can the UNESCO PERSIST Programme Bring about Digital Sustainability through Legacy Software Services?
Janet Delve, David Anderson, Natasa Milic-Frayling, Christopher Lee.......... 316

Workshop 12 Building the Business Case and Funding Advocacy for Data Management Services
Paul Stokes, Neil Beagrie, William Kilbride.......... 318

Workshop 13 Towards Smarter Persistent Identifiers
Jonathan Clark, Juha Hakala, Marcel Ras, Carol Riccalton.......... 320

Tutorial 14 Acquiring and Providing Access to Born-Digital Materials with the BitCurator Environment and BitCurator Access Webtools
Christopher Lee.......... 322

Workshop 15 PREMIS Implementation Fair
Peter McKinney, Eld Zierau, Evelyn McLellan, Angela Dappert.......... 324

Tutorial 16 Relational Database Preservation: Operational Issues and Use Cases
Kuldar Aas, Janet Delve, Krystyna Ohnesorge, Marcel Büchler, Anders Bo Nielsen, Phillip Mike Tømmerholt, Zoltan Lux.......... 326

Workshop 17 Sharing, Using and Re-using Format Assessments
Andrea Goethals, Kate Murray, Michael Day, Kevin L. De Vorsey, Jay Gattuso, Paul Wheatley.......... 328

Workshop 18 Active Data Management Planning: chances and challenges
Felix Engel, Heike Görzig, Rob Baxter, Helen Glaves, Simon Waddington, Matthias Hemmje.......... 330

Index of Authors //.......... 332

PREFACE //

As Programme Co-Chairs, we are delighted that Marie-Christine Doffey, Director of the Swiss National Library, and Elena Balzardi, Vice-director of the Swiss National Library, host and welcome delegates to Bern for the 13th International Conference on Digital Preservation (iPRES), on October 3-6, 2016.

In keeping with previous years, the iPRES 2016 programme is organised into research and practice streams. This format ensures visibility and promotion of both academic research work and the projects and initiatives of institutions involved in digital preservation practices. Furthermore, workshops and tutorials provide opportunities for participants to share information, knowledge and best practices, and explore opportunities for collaboration on new approaches.

Among the highlights of the conference are keynote presentations by distinguished guests and preservation experts: Dr. Robert E. Kahn, Ms Sabine Himmelsbach and Dr. David Bosshart.

Keynotes

Among emerging topics that pre-occupy many are the fast proliferation of digital technologies and the ever increasing production of digital content. These phenomena cause a growing concern about the management of that digital content both for present and future use.

Arguably, the issues of safekeeping digital content that have traditionally stayed within the realm of historical records and memory institutions, are now part of everyday life, posing challenging questions. What happens to our social content? Where will our memories be kept in 50 years? How will the public and scholars, including historians, researchers, and genealogists, know what life was like in the early 21st century? How will we ensure the reproducibility and reuse of scientific output in the future?

Our three keynote presenters address these issues in their specific ways.

Dr. Robert E. Kahn is Chairman, CEO and President of the Corporation for National Research Initiatives (CNRI), which he founded in 1986. He is best known for co-inventing the TCP/IP protocols while working at the U.S. Defense Advanced Research Projects Agency (DARPA) in the early 1970s. TCP and IP are the fundamental communication protocols at the very heart of the Internet. In his recent work, Dr. Kahn has been developing and deploying the concept of Digital Object Architecture. This architecture enables all types of existing information systems to become interoperable, provides for long term persistence of information, and facilitates secure sharing of information. With this in mind, Dr. Kahn reflects on the challenges and opportunities for digital preservation in safeguarding the world's digital heritage. An important challenge for digital preservation is getting agreement on an architecture that can persist in the face of changing technology, thus allowing independent technology choices to be made over time. Another important challenge is to develop and evolve the social structures required for effective management and evolution of such an architecture.

Ms Sabine Himmelsbach is Director of the House of Electronic Arts in Basel. This new home for digital art was established in 2011. Ms Himmelsbach reflects on the complexities of setting up a new organisation dedicated to a new and largely unknown domain - electronic art. That journey starts with the first fundamental question: what is electronic art? Among a broad range of artistic expression, some present conceptual challenges: what is the meaning of a preservation action in the case of 'intentionally ephemeral' art. In essence, there is a tension between the societal need for keeping historical records and the artist's expressed wishes to create art for a moment. Generally, preservation of artists' works involves decisions of how to maintain the context and conceptualisation that underlies the creation of the work, and how to ensure that technical complexity and creative use of technology are understood and maintained over time.

Dr. David Bosshart is CEO of the Gottlieb Duttweiler Institute for economic and social studies in Zurich. Dr. Bosshart's interest is in the impact of technology on society and the changing relationship between humans and machines. He considers the nature and the role of 'digital' in the world today, including the impact on the social, cultural and psychological sense of identity within nations. In that context, it is critical to address the issues facing the long term safekeeping of digital materials that will reflect that identity in the future. That includes the technical complexity, the selection of content to be kept, and the policy and politics of national identity discourse, now and in the future, that may influence the digital preservation agenda.

Programme

The conference programme includes sessions of paper presentations, posters and panels, followed by workshops and tutorials on Wednesday afternoon and Thursday. We received a total of 152 submissions this year and were able to accept 77 of them. The categories of accepted submissions are detailed in the table below.

Submission Type	Accepted	Total
Long papers	19 (63%)	30
Short papers	19 (40%)	47
Panels	3 (33%)	9
Workshops	13 (65%)	20
Tutorials	5 (63%)	8
Posters	18 (47%)	38
Total	77 (51%)	152

The authors had a choice to classify their submission as research focused, practice focused or both. The acceptance rate for research paper submissions was 42% (8 out of 19) and for practitioner paper submissions, 53% (26 out of 49). Papers declaring as both research and practice had an acceptance rate of 50% (4 out of 8) and one paper had no declaration. A few contributions have been withdrawn after acceptance and publication of the programme.

Best Paper

This year's best paper award is sponsored by the Dutch National Coalition on Digital Preservation (NCDD). The Best Paper Award Committee comprised Marcel Ras from NCDD (Chair); Stephen Abrams (California Digital Library); Heike Neuroth (Potsdam University of Applied Science); Libor Coufal (National Library of Australia); and José Borbinha (University of Lisbon).

Here are the three nominees for the best paper award (in order of appearance in the programme):

Will Today's Data Be Here Tomorrow? Measuring The Stewardship Gap by Jeremy York, Myron Gutmann, and Francine Berman

Exhibiting Digital Art via Emulation - Boot-to-Emulator with the EMiL Kiosk System by Dragan Espenschied, Oleg Stobbe, Thomas Liebetraut, and Klaus Rechert

Persistent Web References – Best Practices and New Suggestions by Eld Zierau, Caroline Nyvang, and Thomas Kromann

The winner of the best paper award will be announced during the conference dinner.

Acknowledgments

This year the iPRES conference is hosted by the Swiss National Library and generously supported by the conference sponsors: Swiss National Library, Swiss Federal Archives, CERN, Google, Ex Libris, ETH Bibliothek and SUC P-2 swissuniversities. We extend our warm thanks to all the sponsors, including the Dutch National Coalition on Digital Preservation (NCDD) who provides the best paper award and nestor, the German competence network for digital preservation, who supports the best poster award.

The host Organising Committee, led by the General Chair Barbara Signori, Head of e-Helvetica at the Swiss National Library, is pleased with the community response to the calls for contributions and participation. We wish to express our gratitude to all the Programme Committee members who helped us ensure that iPRES 2016 is a high quality event.

The Programme Co-Chairs would also like to acknowledge the tremendous effort of the local organisers to ensure a smooth running of the conference and a warm welcome to all the attending delegates.

It only remains to wish all the best to the colleagues at the Center for Integrated Area Studies (CIAS), the Kyoto University and the National Institutes for Humanities (NIHU), who will host the iPRES 2017 conference in Kyoto, Japan from September 25-29, 2017. We look forward to seeing you all there.

Nataša Milić-Frayling

Natasa Milic-Frayling
iPRES 2016 Programme Co-Chairs

Steve Knight

CONFERENCE ORGANISATION //

Conference Committee

Hosts
Marie-Christine Doffey, Director, Swiss National Library
Elena Balzardi, Vice-director, Swiss National Library

General Chair
Barbara Signori, Head e-Helvetica, Swiss National Library

Programme Chairs
Natasa Milic-Frayling, Chair in Data Science, University of Nottingham, UK
Steve Knight, Programme Director Preservation Research & Consultancy, National Library of New Zealand
With the support of **Peter McKinney**, Digital Preservation Policy Analyst, National Library of New Zealand

Workshop & Tutorial Chairs
Barbara Sierman, Digital Preservation Manager, Koninklijke Bibliotheek in Den Haag
Raivo Ruusalepp, Head of Development, National Library of Estonia

Poster & Panel Chairs
Andrea Goethals, Manager of Digital Preservation and Repository Services, Harvard Library
Martin Klein, Scientist at Los Alamos National Laboratory, Research Library

Programme Committee

Kuldar Aas	National Archives of Estonia
Stephen Abrams	California Digital Library
Reinhard Altenhöner	State Library Berlin
Christoph Becker	University of Toronto
Jose Borbinha	University of Lisbon
Claire Clivaz	Swiss Institute of Bioinformatics
Libor Coufal	National Library of Australia
Angela Dappert	The British Library
Joy Davidson	University of Glasgow
Kevin De Vorsey	National Archives and Records Administration
Ingrid Dillo	DANS
Ed Fay	University of Southampton
France Fenella	Library of Congress
Rudolf Gschwind	University of Basel
Mariella Guercio	University of Rome Sapienza
Catherine Jones	Science and Technology Facilities Council
Hannes Kulovits	Austrian State Archive
Christopher Lee	University of North Carolina
Michelle Lindlar	German National Library of Science and Technology
Hansueli Locher	Swiss National Library
Peter May	The British Library
Nancy McGovern	MIT Libraries
Salvatore Mele	CERN
Heike Neuroth	University of Applied Sciences Potsdam
Krystyna Ohnesorge	Swiss Federal Archives
Tomasz Parkola	Poznan Supercomputing and Networking Center
Maureen Pennock	The British Library
Buddharaju Raju	National Library Board
Andreas Rauber	Vienna University of Technology
Klaus Rechert	University of Freiburg
Gabriela Redwine	Beinecke Library, Yale University
Thomas Risse	L3S Research Center
João Rocha Da Silva	University of Porto
Lukas Rosenthaler	University of Basel
Seamus Ross	University of Toronto
Sven Schlarb	Austrian Institute of Technology
Jamie Shiers	CERN
Tobias Steinke	German National Library
Armin Straube	nestor
Matthias Töwe	ETH-Bibliothek
Andrew Wilson	University of Portsmouth
Kam Woods	University of North Carolina
Wojciech Wozniak	National Digital Archives
Qu Yunpeng	National Library of China
Zhixiong Zhang	Library of Chinese Academy of Science
Eld Zierau	The Royal Library

Local Organising Committee

Barbara Signori, Head e-Helvetica, Swiss National Library
Martina Speiser, Head of Projects, Organizers Switzerland Ltd.

With the support of Hans-Dieter Amstutz, Maya Bangerter, Genevieve Clavel, Miriam Dubi (Social Media Chair), Doris Egli, Mirjam Gierisch, Stephan Glur, Elisabeth Hermann, Yasmine Keles, Hansueli Locher, Kathrin Marthaler, Nuria Marti, Sujani Ragumar, Fabian Scherler, Simon Schmid, Stephan Schmid, Samantha Weiss, Alena Wenger and Armin Zürcher from the Swiss National Library.

SHORT PAPERS //

A Decade of Preservation: System Migrations in Chronopolis

Sibyl Schaefer
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92092
+1-858-246-0741
sschaefer@ucsd.edu

Mike Smorul
SESYNC
1 Park Place, Suite 300
Annapolis, MD 21401
+1-410-919-4809
msmorul@sesync.org

David Minor
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92092
+1-858-534-5104
dminor@ucsd.edu

Mike Ritter
UMIACS
8223 Paint Branch Drive, Room 3332
College Park, MD 20742
+1-301-405-7092
shake@umiacs.umd.edu

ABSTRACT

This paper provides a historical look at the technical migrations of the Chronopolis digital preservation system over the last ten years. During that time span the service has undergone several software system migrations, moving from middleware-based systems to a suite of individual, finely scoped components which employ widely used and standardized technologies. These transitions have enabled the system to become not only less dependent on interpretation by middleware, but also easier to transfer to new storage components. Additionally, the need for specialized software knowledge is alleviated; any Linux systems administrator should be able to install, configure, and run the software services with minimal guidance. The benefits of moving to a microservices approach have been instrumental in ensuring the longevity of the system through staff and organizational changes.

Keywords

Software Migrations; Microservices; Middleware; iRODS; SRB

1. INTRODUCTION

The Chronopolis system provides long-term, distributed, highly redundant preservation storage. Chronopolis was instituted in 2007 and initially funded by the Library of Congress's National Digital Information Infrastructure and Preservation Program.
A variety of issues related to software and system maintenance and relevant staffing prompted two major software migrations resulting in the service moving from a very centralized, middleware-based system to a set of microservices, defined as "independent but interoperable components that can be freely composed in strategic combinations towards useful ends."[1]

2. CHRONOPOLIS HISTORY

The original Chronopolis partners included the San Diego Supercomputer Center (SDSC) in California, the National Center for Atmospheric Research (NCAR) in Colorado, and the University of Maryland Institute for Advanced Computer Studies (UMIACS) in Maryland. Chronopolis was designed to preserve hundreds of terabytes of digital materials, using the high-speed networks of the partner institutions to distribute copies to each node. In addition to geographical variation, the Chronopolis partner nodes all operate different technology stacks, thus reducing risks associated with specific hardware or software component failures.

Chronopolis has always been administered by people employed within the participating data centers. SDSC provided the original management team, including financials and budgeting, grant management, and HR-related functions. The Center also housed the core system administration staff, who focused on storage systems, software management, and network configurations. NCAR and UMIACS allocated less staff who individually took on broader portfolios. So, for example, a single staff member at NCAR or UMIACS could be responsible for systems, software, networking and code development. These kinds of staffing arrangements grew out of the grant-funded nature of Chronopolis and were appropriate for the network's early development. In subsequent years there have been ongoing efforts to redistribute duties so that some staff positions were fully dedicated to Chronopolis and that these positions were full-time and permanent.

Chronopolis is a founding node in the Digital Preservation Network and also offers preservation storage through the DuraCloud service. Chronopolis was certified as a Trusted Digital Repository by the Center for Research Libraries in 2012 and plans to undergo ISO 16363 certification. Original partner SDSC is no longer an active member of the collaborative, and the University of California San Diego Library has assumed full management authority.

Chronopolis was designed to impose minimal requirements on the data provider; any type or size of digital materials is accepted. Data within Chronopolis are considered "dark." Once ingested, access to the data is restricted to system administrators at each node. These administrators can disseminate a copy of the data stored on their node back to the depositor upon request.

Chronopolis constantly monitors content, especially changes, through the Audit Control Environment (ACE). ACE is a standalone product designed to provide a platform-independent, third party audit of a digital archive. Developed by the ADAPT (A Digital Approach to Preservation Technology) team at UMIACS, research on the ACE software was initially funded by a grant from the National Science Foundation and Library of Congress. Additional development has increased the utility of the program in auditing collections and improved its reporting and logging features.

ACE consists of two components: the Audit Manager and the Integrity Management Service (IMS). The Audit Manager is software that checks local files to ensure they have not been altered. Each Chronopolis node runs the Audit Manager on collections an average of every 45 days. The Integrity Management Service issues tokens used by the Audit Manager to verify that its local store of file digests has not been tampered with. The ADAPT project runs a publically available IMS at ims.umiacs.umd.edu and any group may freely use to register and verify tokens. The Audit Manager software has been released under an open source license and may be downloaded from the ADAPT project website[2].

3. MIGRATIONS

Over the last decade Chronopolis has undergone several infrastructure migrations. Each migration increases the risk of data corruption; ACE has been used as a central piece of the migration process to maintain data integrity.

Two types of migrations have occurred through the Chronopolis lifespan:

1. Standard storage refreshes and upgrades. Storage and network components are generally refreshed every three to five years within the Chronopolis data centers. When Chronopolis was funded primarily through grants, these updates were often coordinated amongst the nodes. Since then, changes have happened asynchronously so that equipment costs are more distributed. Although refreshes and upgrades are major endeavors, these node-internal changes generally do not impact peer nodes other than the upgraded node being temporarily unavailable.
2. Middleware upgrades and changes. Chronopolis has undergone two major software upgrades. The first generation of Chronopolis used the Storage Resource Broker (SRB) to manage data, which was then superseded by the integrated Rule-Oriented Data System (iRODS). Due to a number of factors, in 2014 the Chronopolis sites began migrating out of iRODS and into a homegrown data management system named ChronCore. For the purposes of this paper, we will only be discussing these system transitions and not the more routine storage migrations.

3.1 First Migration: SRB to iRODS

Chronopolis was initially instantiated using the SRB middleware. One motivator for implementing Chronopolis using the SRB was the unified view it provides of different types of storage. During the initial stages of Chronopolis development, both NCAR and SDSC employed a mix of disk and tape storage and the SRB integrated the management of data across both media. This feature diminished in utility as NCAR and SDSC transitioned to large, centrally maintained, disk-based storage pools that were visible to Chronopolis as a single file system.

These new storage pools were directly controlled by iRODS, which was responsible for creating, writing, and reading files on them. UMIACS did not offer a unified file system and was constrained by the total file system size supported by the UMIACS group, so a custom solution, SWAP (Simple Web-Accessible Preservation), was developed. SWAP efficiently mapped files across multiple locally attached disks and servers in a way that required no centralized metadata catalog. Files from this storage were then registered into iRODS post-replication to provide read-only federated access to peer sites. This ensured that future migrations could be performed using standard Unix utilities.

While not evident at inception, SRB's architecture would pose problems for future migration out of the system. All file metadata (names, permission, etc.) were stored in a central metadata catalog while actual file storage was done by renaming the file identifier to this database. This metadata separation posed problems during migration, because the exporting of collection data was only possible using SRB tools, and not at the file system level. This required all sites to store a duplicate copy of all the data in both the old Chronopolis storage and new iRODS locations to ensure that fixity could be checked at all points of data movement. This migration had to occur at each site. Requiring duplicate copies of all data at each node or re-replicating all data between nodes would be a clear constraint on Chronopolis in the future.

3.2 Second Migration: iRODS to ChronCore

Although the federated file system provided an easy means to view contents across the entire Chronopolis network, the administration of iRODS at each site became more of an issue over time, largely due to the dedicated expertise required to maintain the software. The two data centers employing iRODS, NCAR and SDSC, eventually stopped running production iRODS teams, which impacted Chronopolis operations. Additionally, only a small subset of iRODS features was really being applied; previous experience with the SRB made the Chronopolis team wary of technology lock-in so they decided against implementing the sophisticated rule mechanism and metadata capabilities of iRODS in order to facilitate future migrations out of the system. Rather than expending valuable resources on maintaining iRODS support at two nodes, the team decided to migrate to a third system.

ACE was instrumental in moving off of iRODS. Each collection was updated and audited through the REST API to make sure files and tokens were valid. The audit results reported differences between the registered checksums for files and the post-migration captured checksum on local disk, likely due to a bug in the iRODS ACE driver. These discrepancies were resolved by validating the BagIt[3] manifests for each collection and comparing checksums across partner sites. Upon validation that the files were intact, they were removed from ACE and re-registered with accurate fixity information.

4. CHRONCORE

The main purpose of ChronCore is to package and distribute data securely throughout the system, providing several levels of bit auditing to ensure that nothing is lost in transmission. The distributed architecture of Chronopolis led to the creation of distributed services. As each core service emerged, it was assigned scoped operations depending on its place in the Chronopolis pipeline. ChronCore consists of three such scoped services: intake, ingest, and replication. Currently only the UCSD library node runs the intake and ingest services, which package, record, and stage data for replication. All partner sites run the replication services, which poll the ingest service hourly to determine if new collections have been staged for replication.

4.1 ChronCore Services

4.1.1 Intake

Content is submitted by a depositor through one of the Chronopolis Intake services. If the content is bagged and a manifest is present, the Intake service will verify the manifest and, if valid, register the collection with the Ingest server. If the content has not been previously packaged, the Intake service will bag the content before registering it with the Ingest server.

4.1.2 Ingest

The Ingest service serves as a central registry for content in Chronopolis. It generates ACE tokens, which provide provenance data about when content was first ingested and what fixity values it arrived with. Once tokenization is complete, the Ingest service will create replication requests which are picked up by each partner site. Replication of both the content and tokens are served through a RESTful API.

4.1.3 Replication

Each partner site runs a Replication service that periodically queries the Ingest service API and performs replications on any requests. The general flow of events for a replication is:

1. Query for new requests.
2. Transfer data (rsync/https/gridftp).
3. Respond with checksum of transferred content.
4. If content is valid, register it with the local Audit Manager and kick off the initial local audit.
5. Close transaction.

If a replication fails, the Ingest server is notified and a new request needs to be generated with the replication server. The cause of failure is first manually reviewed to determine if the cause was intermittent (network issues) or something more serious (bit flips).

4.2 Industry Standard Technologies

As a lightweight system of microservices, ChronCore does not contain the entire breadth of functionality that the previously employed middleware systems offered; time has proven that this advanced functionality is not necessary for Chronopolis operations. Instead of developing new tools or implementing new technologies, project leaders decided to take advantage of older, simpler technologies that have been demonstrated over time to operate at the necessary scales.

- SSH: by providing access to a service account at each site, a federated system can be 'mocked' with strict access controls ensuring no data is tampered with.
- rsync: this is a proven transport mechanism for transferring data to each site. It allows for transfers to be restarted, mitigating the impact of intermittent network problems. Over Chronopolis' lifetime, the community at large has shown that this tool could scale to Chronopolis-sized connections.
- HTTP/REST: REST/JSON has rapidly become an accepted communication protocol, replacing older vendor-specific binary protocols. In addition, its support by numerous languages and toolkits assures vendor or language lock-in will not be an issue.

5. CONCLUSIONS

Throughout the last ten years, Chronopolis has been able to migrate a set of replicated collections through a variety of systems while maintaining bit integrity and provenance. The experience gained from system migrations has led the Chronopolis team to espouse the following tenets.

Use Independent Microservices. Chronopolis has migrated from very centralized middleware systems to a mix of off-the-shelf and custom technologies. By creating a system of specialized services, each can follow its natural technical lifecycle and be replaced as appropriate. ACE was an early example of this model and has persisted as a key part of the Chronopolis infrastructure even as every other component has been switched out.

Always have direct data access. The move from systems where raw file data was hidden below layers of middleware to standard file system storage has removed a reliance on specialized tools or software and enabled the use of more widely used and supported utilities. Conceptually this also follows the independent services lesson mentioned previously, as it is a critical aspect in allowing technologies to be switched out as necessary. In previous implementations, access to files was dependent on metadata services managed by the middleware system. Potential loss of this metadata catalog due to a higher-level service failure created increased risk within the system, requiring additional care to ensure the catalog was highly available and recoverable. Complete loss of these services could render the on-disk data unusable. Information about provenance, preservation actions, and even original filenames could also be lost. These middleware systems also required each Chronopolis partner to maintain in-house expertise to support this custom software. Maintaining the necessary staff expertise at all three sites increased the operational costs of the network.

Choose "boring" technologies that don't require specialized expertise[4]. Chronopolis has changed not only the software it uses over time but also the staff that runs the system. Each node has experienced significant staff turnover over the past ten years; sometimes within as little as a year one or more nodes would undergo changes in management. By migrating from large proprietary systems to common technologies, Chronopolis has greatly increased its resilience to personnel changes at any of its sites. All of the core tools are well supported, have large, active user communities, and are within the skill sets of most system administrators. Should there be a personnel shortage at a site, it would be fairly easy to contract the necessary expertise to keep the node up and running. Using widely adopted tools also lowers the barrier for new nodes to participate in Chronopolis, and was instrumental in the ease with which the management of the San Diego node transferred from SDSC to the UCSD Libraries.

6. REFERENCES

[1] Abrams, S., Cruse, P., Kunze, J., and Minor, D. 2010. Curation Micro-services: A Pipeline Metaphor for Repositories. *Journal of Digital Information*. 12, 2. (April. 2011), https://journals.tdl.org/jodi/index.php/jodi/article/view/1605/1766.

[2] ADAPT: An Approach to Digital Archiving and Preservation Technology. https://adapt.umiacs.umd.edu.

[3] Kunze, J. et al. The BagIt File Packaging Format (VO. 97) Network Working Group Internet-Draft. https://tools.ietf.org/html/draft-kunze-bagit-13

[4] McKinley, Dan. Choose Boring Technology. http://mcfunley.com/choose-boring-technology

Digital Preservation through Digital Sustainability

Matthias Stuermer
Institute of Information Systems at
University of Bern
Engehaldenstrasse 8, 3012 Bern
+41 76 368 81 65
matthias.stuermer@iwi.unibe.ch

Gabriel Abu-Tayeh
Institute of Information Systems at
University of Bern
Engehaldenstrasse 8, 3012 Bern
+41 76 505 00 55
gabriel.abu-tayeh@iwi.unibe.ch

ABSTRACT

The concept of digital sustainability introduces a holistic approach on how to maximize the benefits of digital resources for our society. The nine basic conditions for digital sustainability also provide a contribution to potential solutions to the challenges of digital preservation. Elaborateness, transparent structures, semantic data, distributed location, an open licensing regime, shared tacit knowledge, participatory culture, good governance, and diversified funding support the long-term availability of digital knowledge. Therefore, in this conceptual paper, we explain the links between digital sustainability and digital preservation in order to increase the impact of both. We conclude by presenting the political agenda of the Swiss parliamentary group for digital sustainability.

Keywords

Digital sustainability, digital preservation, open source software, open data, open standards, linked open data

1. INTRODUCTION

The discussion on sustainable development started at a global level in 1987 when the United Nation's World Commission on Environment and Development, led by Gro Harlem Brundtland, published the report "Our Common Future" [59]. Today, sustainable development represents a vision more relevant than ever, perhaps the most prominent example being the United Nations' Sustainable Development Goals, launched in 2015 [45] [55].

Most literature on sustainable development focuses on natural resources, human rights, and economic development. However, more recently, sustainability has also become a topic in digital preservation, software engineering, and information systems research. For example, the Blue Ribbon Task Force on Sustainable Digital Preservation and Access in 2010 presented a comprehensive report on the economic challenges of providing sustainable access to digital information [49]. Maintenance of software is hindered because of technical debt of its architecture leading to the insight that sustainability of software systems is important for their resilience, adaptability, and durability [3]. Therefore, several software engineering researchers have recently released a 0.5 version of their Karlskrona Manifesto for Sustainability Design of software [4].

Our holistic notion of digital sustainability covers digital information as well as software systems. The initial idea was briefly introduced in a recent conference publication [51]. An in-depth conceptual working paper derives the nine basic conditions for digital sustainability from sustainability studies, knowledge management, digital information, and innovation literature [52].

In the following article, we link the nine conditions for digital sustainability with examples from the field of digital preservation since this presents a highly relevant stream of research for our knowledge society.

2. BASIC CONDITIONS FOR DIGITAL SUSTAINABILITY

The basic conditions for digital sustainability include legal, technical, organizational, and financial requirements we consider necessary for the creation and use of sustainable digital artifacts. While the first four conditions address the digital artifact itself, the latter five target the surrounding ecosystem (Table 1). This illustrates an important aspect of the concept of digital sustainability: We find that it is not only the characteristics of digital resources that are relevant for its sustainability, but also the community of people and organizations involved in the digital resource. It is therefore essential for our concept of digital sustainability on the one hand that suitable properties of the digital asset are ensured, while on the other hand maintaining a sound ecosystem that continuously updates and grows the digital artifact.

Table 1: Basic conditions for digital sustainability

Conditions regarding the digital artifact:	1	Elaborateness
	2	Transparent structures
	3	Semantic data
	4	Distributed location
Conditions regarding the ecosystem:	5	Open licensing regime
	6	Shared tacit knowledge
	7	Participatory culture
	8	Good governance
	9	Diversified funding

2.1 Elaborateness

Digital resources create immediate as well as long-term value to society through their elaborateness. For instance, data quality requires characteristics such as accuracy, relevancy, timeliness, completeness and many more characteristics [57]. Within software development, modularity of the source code is crucial. If the code is modular it can easily be enhanced by programmers because it is not necessary to completely understand the source code in order to improve and enhance it [30].

Quality of data plays a significant role within digital preservation. On the one hand, libraries are often confronted with errors in documents and their metadata [2]. Within the documents there are often typographical errors, scanning and data conversion errors, as well as 'find and replace' errors. Metadata quality is defined by characteristics such as accuracy, comprehensiveness, consistency, flexibility and many more, some obviously with competing properties [33]. The growing volume and digitization quality of digital assets, steadily increasing the demands for data storage, pose another challenge in preserving data quality [13]. While, in the early days, preservation targeted full information capture of media by archiving microfilm and alkaline paper, today technology facilitates the digitization of analog material in a high quality. Therefore, preserving data quality is also a question of financial resources [54].

2.2 Transparent Structures

In addition to the elaborateness of a digital artifact, its technical openness of content and software is essential for digital sustainability. Digital artifacts can be best used and further developed if their inner structures are transparent and well-documented. For example, access to the source code facilitates the re-use of open source components, saving substantial development costs [21]. Alternatively, open standards such as the Open Document Format (ODF) are developed through a participatory process within a standardization body (in the case of ODF, the "Organization for the Advancement of Structured Information Standards", OASIS [8]), fully documented and made publicly available, as well as being integrated into various software tools such as LibreOffice [19]. An open standard allows the development of software implementing the standard, grants low-cost (or even free), universal access to the standards, and assures that the standard has been developed using a participatory approach [14]. The architectural transparency of software and content thus allows verification by any technically skilled person, thereby reducing errors and increasing trust in digital artifacts. Therefore, transparent structures are another basic condition for digital sustainability.

Open standards and open file formats are particularly important for digital preservation. While there are various definitions and lists of criteria characterizing long-term preservation formats, all of these include "open specification", "transparency", or "openness" as one of their requirements [38]. Researchers on digital preservation thus agree that "open standard" is a crucial criteria for any content to be made long-term accessible [56].

However, having the data in an open format is but one side of the coin. Appropriate software is always necessary to read the documents. While some file formats are pretty straightforward to read (e.g. plain text) other content, such as structured documents, images, video, music, or Geographical Information Systems (GIS,) is stored in highly complex standards. Their technical specifications might be openly available, as is the case with the Microsoft document standard OOXML. However, the extensive documentation of such standards (OOXML specification is approximately 7000 pages [8]) indicates the effort required to implement such a file format. Only a few - if not only a single corporation (the one who has drafted the specification) - will be able to program an implementation, often resulting in proprietary software. Those software products become an object of control for a single company, thus decreasing the sustainability of development. Therefore, the availability of an open source implementation of an open standard is required to support a file format in the long term.

2.3 Semantic Data

In order to make the vast amount of digital resources accessible from an information management perspective, it is highly beneficial to enrich the data with metadata [24]. Structured semantic data makes complex digital artifacts machine-readable [7] and also more easily comprehensible to humans by adding meaningful information about the data [18]. Various semantic platforms such as DBpedia [1] [7] and Wikidata [53] [49] have emerged in recent years, providing knowledge graphs in order to make large volumes of digital information accessible to humans and machines.

Within the digital preservation literature, for example, the Digital Preservation Recommender (DiPRec) system [20] addresses the issue of structured information of digital assets through Linked Open Data (LOD). This approach applies the semantic Web and linked open data paradigms in order to "transform the web from a pool of information into a valuable knowledge source of data". The importance of metadata for records keeping was already pointed out by the Open Archival Information System (OAIS) reference model [28] and the ISO standard 16363 on "Audit and Certification of Trustworthy Digital Repositories". They both provide thorough conceptual guidance on sustainability of digital preservation systems.

2.4 Distributed Location

The redundant storage of information in different locations decreases the risk of it being lost as a result of hardware crash or other accidents. Ideally, digital resources are replicated and stored in a decentralized way through peer-to-peer technology like the Bitcoin Blockchain [43] [15] in order to maximize independence from any single storage provider.

Within digital preservation, institutional repositories enable educational organizations to provide access to assets of an institution, such as research results and educational resources. However, the long-term availability of the service is a challenge, as continuity depends on the way the information systems are managed by the particular institution [23]. A successful approach was introduced in the beginning of 2000 when David Rosenthal and Vicky Reich launched LOCKSS (Lots Of Copies Keep Stuff Safe) at the Stanford University Libraries [39]. Since these early days of the Internet this open source platform provides continued access to scientific journals based on peer-to-peer technology [44].

2.5 Open Licensing Regime

As explained above, in addition to the conditions of digital artifacts, there are essential properties of its ecosystem that ultimately influence digital sustainability. Part of this is the legal framework playing a crucial role for digital artifacts.

Text, images or software are by default protected by intellectual property rights [40]. While this mechanism is the basis for many business models, it hinders the use of these digital assets and thus decreases their potential for society as a whole. Only if content or source code is explicitly published under an open license – such as the Creative Commons [26] or an open source license [47] [46] – is that digital resource available to all without restriction. The notion of legal public release of digital assets dates back to the 1980's when Richard M. Stallman drafted the GNU General Public License [50]. About two decades later, this principle of freely available digital assets was transferred to content such as open educational resources [10] and open data [32]. A generalized definition of 'open' is provided by the Open Definition, which states "Open means anyone can freely access, use, modify, and share for any purpose" [1]. An open licensing regime enables the unrestricted use and modification of digital assets and thus forms a basic condition for digital sustainability.

As far as digital preservation is concerned, open licenses are highly practical for the storage of digital materials e.g. by libraries. Usually, there is a conflict of interest between the copyright holder, such as the publisher and e.g. the national library charged by the public to preserve digital heritage [27]. While there are certain circumventions, such as the "fair use" approach [17], cultural heritage institutions benefit substantially if scientific output is published under an open access regime [31]. This resolves all intellectual property restrictions by granting long-term access without any legal limitations.

2.6 Shared Tacit Knowledge

Using and enhancing digital resources requires specific skills and experiences on how to interpret and modify the technical structures. In knowledge management theory, such uncodified experience is called 'tacit knowledge' and enables individuals and groups to understand and apply technologies and create further knowledge [35]. While digital resources do not diminish through usage, they do need to be updated and adapted continuously to reflect the changing environment. Thus, knowledge about certain technologies is best preserved through collective intelligence [6], meaning tacit knowledge about the digital resource should be spread as widely as possible.

Making digital resources available long-term requires skills and knowledge on how to properly handle them and correctly interpret the stored information. Therefore, not only the explicit forms - such as recorded data - are necessary for digital sustainability; tacit knowledge is also crucial to be able to maintain and interpret the resources in the long-term. Digital preservation scholars have identified problems when tacit knowledge is lost, including an increased risk of not being able to read and understand the data in the future [48]. This illustrates the critical role of such uncodified knowledge.

2.7 Participatory Culture

Assuming knowledge is being shared among various stakeholders, how should sustainable digital resources be developed further? Experience from open source projects (Linux kernel etc.) or open content communities (Wikipedia etc.) have shown that an active ecosystem leads to significant contributions from outsiders such as volunteers [37] and corporations [58]. Such dispersed communities gather the expertise from an international set of contributors, ideally leading to high-quality peer-reviewed processes of knowledge creation.

Archives and other digital heritage institutions have the potential to benefit greatly from these kinds of crowdsourcing methods. Quality assurance and information gathering processes, as well as assessments, have been testing a number of participatory patterns [12]. In addition, crowdsourcing projects promoted by galleries, libraries, archives, museums, and educational institutions have started to be applied, leading to positive results and empirical insights [9]. For instance, the Brooklyn Museum and other GLAM (galleries, libraries, archives, and museums) institutions made successful experiments with crowdsourcing games (Games with a Purpose, GWAP) where citizens conducted microtasks such as tagging content and validating data [42].

2.8 Good Governance

Nowadays, many digital resources are produced and controlled by corporations. However, centralized control by a single entity might not be an appropriate governance basis for a sustainable digital resource as it becomes directly linked to the organization's continuity. While technology companies and innovative business models are considered part of sustainable digital resources [53], they should remain independent from self-serving commercial interests and control in the hands of only a few individuals. Open source projects integrate the possibility of 'forking', signifying the division of the developer community [36]. Although such events can bring turmoil and wastage of resources, they are a crucial element within open source communities, potentially leading to more sustainable governance structures and more effective collaboration [19]. Thus, good governance among contributors and other stakeholders represents another condition of sustainable digital resources.

In digital preservation projects decisions, often need to be taken on which information is digitalized and made available publicly and which is not [27]. Not all data can be digitally published since the resources of archives are limited and certain sources could result in too much effort. Therefore, publication should follow a careful planning and decision-making process including all relevant stakeholders. Ideally, the selection procedure leads to "well-documented, well-argued and transparent decisions" [5]. Another example indicates the importance of widely supported governance structures: In 2003, the UNESCO Charter acknowledged that digital heritage is essential for our society [29]. Within the charter, a multilevel approach was proposed: Universal strategies, strategies adapted to geographical and national configurations and the involvement of authors, publishers and other relevant stakeholders are required. The development of cultural heritage should not be based on a selection made by a single institution.

2.9 Diversified Funding

While governance may be shared broadly among various stakeholders, ultimately, it tends to be financial capabilities that direct the use of resources. Therefore, diversified funding reduces control by a single organization, thus increasing the independence of the endeavor. There are a variety of funding models available, as explained with the example of open educational resources [16]: the endowment model (interests paid), the membership model (all interested organizations pay a fee), the donations model (voluntary donations), the conversion model (selling of added value services), the contributor-pay model (contributors are charged), the sponsorship model (public relations by corporations), the institutional model (a public institution pays), the government model (a government agency pays), and the partnership and exchanges model (costs are split among various institutions).

Digital heritage work is for the most part funded by public institutions or by some other not-for-profit sources, such as lottery funds [41]. As such, it is presumed to be less prone to commercial exploitation by corporations. Nevertheless, diversified funding of digital preservation projects supports scientific independence and increases public awareness of the societal impact of digital heritage. In order to leverage public funding, incentives should be introduced to motivate private investments into digital preservation activities [11].

3. POLITICAL AGENDA FOR DIGITAL SUSTAINABILITY

As with many initiatives relating to sustainable development, most people might agree upon the goals. However, the question remains how these aims can be implemented successfully.

One approach addresses the policy level in order to advance the attainment of targets relating to digital sustainability. In Switzerland, there is a national parliamentary group lobbying for the concept of digital sustainability[1]. The group was founded in 2009 in order to increase the creation and use of open standards, open source software, open content, open data, and open access [34] [22].

Among others, this nonpartisan group of parliamentarians advocates the following issues regarding digital sustainability:

Public funding of digital resources should follow the conditions for digital sustainability. Thus, institutions like the national archive should not only ensure that digital heritage data is stored within open formats, but also that the requisite software is available under free licenses, such as open source software.

Public institutions should prioritize the procurement of open source software. In order to decrease dependencies from proprietary software vendors, public tenders should favor bids offering open source software solutions. Libraries in particular are not yet fully exploiting the potential of open source software, as academics already noted as long ago as 2007 [25].

Research funding should focus on open science principles. Publicly funded research should provide the aggregated results in open access journals and the research data as open data. Furthermore, all software developed during research activities should be published as open source.

Political statements and policy interventions, like the ones outlined above, are helping to promote digital sustainability in the public sector, thereby advancing the notion for digital preservation also.

[1] www.digitale-nachhaltigkeit.ch

4. REFERENCES

[1] Bates, J. 2012. "This is what modern deregulation looks like" : co-optation and contestation in the shaping of the UK's Open Government Data Initiative. *The Journal of Community Informatics*. 8, 2 (Feb. 2012).

[2] Beall, J. 2006. Metadata and data quality problems in the digital library. *Journal of Digital Information*. 6, 3 (2006).

[3] Becker, C. 2014. Sustainability and longevity: Two sides of the same quality? *mental*. 20, (2014), 21.

[4] Becker, C., Chitchyan, R., Duboc, L., Easterbrook, S., Penzenstadler, B., Seyff, N. and Venters, C.C. 2015. Sustainability design and software: The karlskrona manifesto. *Proceedings of the 37th International Conference on Software Engineering-Volume 2* (2015), 467–476.

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

[6] Benkler, Y., Shaw, A. and Hill, B.M. 2015. Peer Production: A Form of Collective Intelligence. *Handbook of Collective Intelligence*. (2015), 175.

[7] Berners-Lee, T. and Hendler, J. 2001. Publishing on the semantic web. *Nature*. 410, 6832 (Apr. 2001), 1023–1024.

[8] Blind, K. 2011. An economic analysis of standards competition: The example of the ISO ODF and OOXML standards. *Telecommunications Policy*. 35, 4 (Mai 2011), 373–381.

[9] Carletti, L., Giannachi, G., Price, D., McAuley, D. and Benford, S. 2013. Digital humanities and crowdsourcing: an exploration. (2013).

[10] Caswell, T., Henson, S., Jensen, M. and Wiley, D. 2008. Open Content and Open Educational Resources: Enabling universal education. *The International Review of Research in Open and Distributed Learning*. 9, 1 (Feb. 2008).

[11] Chowdhury, G.G. 2015. How to improve the sustainability of digital libraries and information Services? *Journal of the Association for Information Science and Technology*. (Nov. 2015), n/a-n/a.

[12] Clough, P., Sanderson, M., Tang, J., Gollins, T. and Warner, A. 2013. Examining the Limits of Crowdsourcing for Relevance Assessment. *IEEE Internet Computing*. 17, 4 (Jul. 2013), 32–38.

[13] Conway, P. 2010. Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas. *The Library Quarterly*. 80, 1 (Jan. 2010), 61–79.

[14] Coyle, K. 2002. Open source, open standards. *Information Technology and Libraries*. 21, 1 (2002), 33.

[15] De Filippi, P. 2014. Bitcoin: a regulatory nightmare to a libertarian dream. *Internet Policy Review*. 3, 2 (2014).

[16] Downes, S. 2007. Models for sustainable open educational resources. (2007).

[17] Fisher III, W.W. 1988. Reconstructing the Fair Use Doctrine. *Harvard Law Review*. 101, 8 (Jun. 1988), 1659.

[18] Floridi, L. 2005. Is Semantic Information Meaningful Data? *Philosophy and Phenomenological Research*. 70, 2 (März 2005), 351–370.

[19] Gamalielsson, J. and Lundell, B. 2014. Sustainability of Open Source software communities beyond a fork: How and why has the LibreOffice project evolved? *Journal of Systems and Software*. 89, (Mar. 2014), 128–145.

[20] Gordea, S., Lindley, A. and Graf, R. 2011. Computing recommendations for long term data accessibility basing on open knowledge and linked data. *Joint proceedings of the RecSys*. (2011), 51–58.

[21] Haefliger, S., von Krogh, G. and Spaeth, S. 2008. Code Reuse in Open Source Software. *Management Science*. 54, 1 (Jan. 2008), 180–193.

[22] Hillenius, G. 2009. CH: Parliamentarians begin group on digital sustainability | Joinup.

[23] Hockx-Yu, H. 2006. Digital preservation in the context of institutional repositories. *Program*. 40, 3 (2006), 232–243.

[24] Jackendoff, R.S. 1990. *Semantic Structures*. Cambridge: MIT Press.

[25] Jaffe, L.D. and Careaga, G. 2007. Standing up for open source. *Library Philosophy and Practice*. 9, 2 (2007), 21.

[26] Katz, Z. 2005. Pitfalls of Open Licensing: An Analysis of Creative Commons Licensing. *IDEA: The Intellectual Property Law Review*. 46, (2006 2005), 391.

[27] Lavoie, B. and Dempsey, L. 2004. Thirteen ways of looking at... digital preservation. *D-Lib magazine*. 10, 7/8 (2004), 20.

[28] Lee, C.A. 2010. Open archival information system (OAIS) reference model. *Encyclopedia of Library and Information Sciences*. (2010), 4020–4030.

[29] Lusenet, Y. de 2007. Tending the Garden or Harvesting the Fields: Digital Preservation and the UNESCO Charter on the Preservation of the Digital Heritage. *Library Trends*. 56, 1 (2007), 164–182.

[30] MacCormack, A., Rusnak, J. and Baldwin, C.Y. 2006. Exploring the Structure of Complex Software Designs: An Empirical Study of Open Source and Proprietary Code. *Management Science*. 52, 7 (Jul. 2006), 1015–1030.

[31] McCray, A.T. and Gallagher, M.E. 2001. Principles for digital library development. *Communications of the ACM*. 44, 5 (2001), 48–54.

[32] Miller, P., Styles, R. and Heath, T. 2008. Open Data Commons, a License for Open Data. *LDOW*. 369, (2008).

[33] Moen, W.E., Stewart, E.L. and McClure, C.R. 1998. Assessing metadata quality: Findings and methodological considerations from an evaluation of the US Government information locator service (GILS). *Research and Technology Advances in Digital Libraries, 1998. ADL 98. Proceedings. IEEE International Forum on* (1998), 246–255.

[34] Neuroni, A.C., Riedl, R. and Brugger, J. 2013. Swiss Executive Authorities on Open Government Data -- Policy Making beyond Transparency and Participation. (Jan. 2013), 1911–1920.

[35] Nonaka, I. and Konno, N. 1998. The concept of "ba": Building a foundation for knowledge creation. *California management review*. 40, 3 (1998), 40–54.

[36] Nyman, L. and Lindman, J. 2013. Code Forking, Governance, and Sustainability in Open Source Software. *Technology Innovation Management Review*. January 2013: Open Source Sustainability (2013), 7–12.

[37] O'Mahony, S. 2007. The governance of open source initiatives: what does it mean to be community managed? *Journal of Management & Governance*. 11, 2 (Jun. 2007), 139–150.

[38] Park, E.G. and Oh, S. 2012. Examining attributes of open standard file formats for long-term preservation and open access. *Information Technology and Libraries*. 31, 4 (2012), 46–67.

[39] Reich, V. and Rosenthal, D.S.H. 2001. LOCKSS: A Permanent Web Publishing and Access System. *D-Lib Magazine*. 7, 6 (Jun. 2001).

[40] Reichman, J.H. 1995. Universal Minimum Standards of Intellectual Property Protection under the TRIPS Component of the WTO Agreement. *The International Lawyer*. 29, 2 (1995), 345–388.

[41] Richardson, L. 2013. A Digital Public Archaeology? *Papers from the Institute of Archaeology*. 23, 1 (Aug. 2013), 10.

[42] Ridge, M. 2013. From Tagging to Theorizing: Deepening Engagement with Cultural Heritage through Crowdsourcing. *Curator: The Museum Journal*. 56, 4 (Oktober 2013), 435–450.

[43] Ron, D. and Shamir, A. 2013. Quantitative analysis of the full bitcoin transaction graph. *Financial Cryptography and Data Security*. Springer. 6–24.

[44] Rosenthal, D.S.H., Vargas, D.L., Lipkis, T.A. and Griffin, C.T. 2015. Enhancing the LOCKSS Digital Preservation Technology. *D-Lib Magazine*. 21, 9/10 (Sep. 2015).

[45] Sachs, J.D. 2012. From Millennium Development Goals to Sustainable Development Goals. *The Lancet*. 379, 9832 (Jun. 2012), 2206–2211.

[46] Scacchi, W. and Alspaugh, T.A. 2012. Understanding the role of licenses and evolution in open architecture software ecosystems. *Journal of Systems and Software*. 85, 7 (Jul. 2012), 1479–1494.

[47] Sen, R., Subramaniam, C. and Nelson, M.L. 2011. Open source software licenses: Strong-copyleft, non-copyleft, or somewhere in between? *Decision Support Systems*. 52, 1 (Dezember 2011), 199–206.

[48] Smit, E., Van Der Hoeven, J. and Giaretta, D. 2011. Avoiding a Digital Dark Age for data: why publishers should care about digital preservation. *Learned Publishing*. 24, 1 (Jan. 2011), 35–49.

[49] Smith Rumsey, A. 2010. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*. Blue Ribbon Task Force on Sustainable Digital Preservation and Access.

[50] Stallman, R. 2002. *Free software, free society: Selected essays of Richard M. Stallman*. Lulu. com.

[51] Stuermer, M. 2014. Characteristics of digital sustainability. *Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance* (2014), 494–495.

[52] Stuermer, M., Abu-Tayeh, G. and Myrach, T. 2016. Digital sustainability: Basic conditions of digital artifacts and their ecosystem. *Working paper, University of Bern*. (2016).

[53] Stuermer, M., Spaeth, S. and Von Krogh, G. 2009. Extending private-collective innovation: a case study. *R&D Management*. 39, 2 (Mar. 2009), 170–191.

[54] Thibodeau, K. 2002. Overview of technological approaches to digital preservation and challenges in coming years. *The state of digital preservation: an international perspective*. (2002), 4–31.

[55] United Nations 2015. *Transforming our world: the 2030 Agenda for Sustainable Development*. United Nations.

[56] Vilbrandt, C., Pasko, G., Pasko, A., Fayolle, P.-A., Vilbrandt, T., Goodwin, J.R., Goodwin, J.M. and Kunii, T.L. 2004. Cultural heritage preservation using constructive shape modeling. *Computer Graphics Forum* (2004), 25–41.

[57] Wang, R.Y. and Strong, D.M. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*. 12, 4 (März 1996), 5–33.

[58] West, J. and O'Mahony, S. 2008. The Role of Participation Architecture in Growing Sponsored Open Source Communities. *Industry & Innovation*. 15, 2 (Apr. 2008), 145–168.

[59] World Commission on Environment and Development 1987. Report of the World Commission on Environment and Development: Our Common Future. (1987).

Building a Future for our Digital Memory
A National Approach to Digital Preservation in The Netherlands

Marcel Ras
National Coalition on Digital Preservation
Prins Willem-Alexanderhof 5
2509 LK Den Haag
+31 6 147 776 71
marcel.ras@ncdd.nl

ABSTRACT

In 2015 the national Network for Digital Heritage was established. This network is based on three pillars: to make the digital heritage wider visible, better usable and more sustainable preserved. A series of collaborative projects are in progress since Summer 2015, framed within three working programs, all with their own but integrated set of dedicated actions in order to create a national infrastructure in the Netherlands, based on an optimal use of existing facilities. In this paper the focus is on the activities related to the sustainable preservation of the Dutch national digital heritage. What are the developments and where are we now, with the program running for a year and the first results are delivered.

Keywords

Digital Preservation, Dutch Coalition on Digital Preservation, national infrastructure, collaboration, national collaboration

1. INTRODUCTION

Collaboration in digital preservation has a long-standing tradition as partners within the same domain are working together for a long time (libraries, archives, data centers). Facing the rapid technological developments, the growing amount of digital material and the growing complexity of digital objects, it seems clear that no one institution can do digital preservation on its own. So close collaboration between the organizations involved in digital preservation is required. And organizations were aware of this very early.

Collaborations had a firm basis in research and development issues and were framed within large-scale national and international projects delivering usable results for organizations. An additional deliverable of these intensive projects was the growth of a common understanding of each other's issues and positions. You could say that we learned to know each other much better then before.

In 2002, the DPC [1] was founded as a "collaborative effort to get digital preservation on the agenda of key decision-makers and funders". Similar intentions led to the foundation of nestor [2], NCDD [3], and NDSA [4]. OPF [5] and Presto Center [6] were set up as international competence centers. Overall, these organizations serve as platforms for training, knowledge exchange and the study of specific preservation related issues.

Examples of collaborative efforts were already presented at previous conferences. At the 2014 iPRES conference the national digital repository of Ireland was discussed [7]. Darryl Mead from the National Library of Scotland described the effort to create a national preservation infrastructure in Scotland [8]. And in Finland the national library, archives and museums share already an infrastructure. [9] This development is also reflected in Recommendation 3 in the Roadmap of the European 4C Project (Collaboration to Clarify the Costs of Curation), stating, "Develop scalable services and infrastructure", with the explicit benefit of enabling "the realization of further cost reductions by improving efficiency of the workflows necessary to undertake digital curation" [10]. And finally, at the 2015 iPRES conference the first steps in creating a national infrastructure for digital preservation in The Netherlands was presented [11].

2. DUTCH COALITION ON DIGITAL PRESERVATION (NCDD)

On May 21st 2007 a group of organisations took the initiative to set up a coalition to address the problem of digital preservation in The Netherland in a collaborative way. This coalition of the willing became a foundation in 2008 with its mission to establish an infrastructure (organisational and technical) to guarantee long-term access to digital information in The Netherlands. NCDD acts as the national platform for exchange of knowledge and expertise and has a role in coordinating and facilitating the establishment of a national network in which long term access to digital information which is of crucial importance for science, culture and society is guaranteed.

Cross-domain collaboration and agreement are key to realizing high-quality, effective and efficient digital information management. The NCDD partners are advancing this collaborative approach by searching for the best solutions across the board of the public domain. This explicitly includes the interests of smaller organizations which, due to a lack of technical facilities, organization and knowledge, are not capable of ensuring reliable digital management on their own. In 2013 NCDD made it part of her strategy to work on this collaborative model that should result in a distributed national infrastructure.

Following on a national survey [12], the NCDD in 2010 formulated a strategic agenda [13]. This agenda consisted of a description of the major steps to be taken on a national level in the Netherlands in order to address the issues described in the survey. The strategy is centred on four themes: (1) knowledge-sharing; (2) development of a scalable and usable infrastructure for long-term management of digital information; (3) cost management; and (4) development of co-ordination in collection development policies. NCDD partners are working on realizing these themes by conducting collaborative projects. Project teams are made up of experts from various organizations (coalition members as well as other collection managing institutions) and are led by a representative of one of the NCDD partners. In this way, we can pool our resources and expertise to expand our knowledge and attain shared solutions.

It was also thought necessary to create a sense of urgency towards policy makers on all levels, with the message that we had to act, and act on a national level, to ensure long-term access of digital information. Within the sense of urgency the focal point was the development towards a national infrastructure. Therefore NCDD and especially the partners within the NCDD took the lead in addressing the problem on a policy level, but also on a practical level. It was decided that under the umbrella of the NCDD coalition, the large heritage institutes in The Netherlands would work out a "collaborative model", setting up collaborative facilities or share facilities where possible. Which in reality would not always be the case.

The first series of NCDD projects started in 2014 [14]. Apart from the collaborative projects, the NCDD carried out a survey into a national infrastructure for sustained access to digital information, which was commissioned and financed by the Ministry of Education, Culture and Science [15]. The results of this investigation, combined with the collaborative projects, are the puzzle pieces from which this national infrastructure is to be created. They effectively realized first results of the goals set out in the NCDD's strategic agenda. The next steps will be worked out in the Work program three of the NDE (Preservable Digital Heritage), where the current situation will be turned into a networked future.

3. DIGITAL PRESERVATION AS A NATIONAL PROGRAM

The objective of this Work Program is to create, through cross-domain collaboration, a shared infrastructure that guarantees sustainable access to digital information. The assumption is that this cooperation will lead to an increased effectiveness, greater efficiency and cost reductions. Most of the activities in this work program have been started and scheduled within the NCDD strategic agenda.

Initiated by the Ministry of Education, Culture and Science, the Digital Heritage Network (NDE) was set up in 2014. This network consists of a number of large organizations occupying key positions in the field of digital heritage, including the NCDD partners. Together, these organizations aim to improve the visibility, usability and sustainability of digital heritage materials from every domain. To this end, the Digital Heritage Network has developed a three-pronged strategy covering Visible, Usable and Sustainable Digital Heritage, respectively. A work package has been established for each of these aspects, outlining the projects necessary to achieve its central goals [16]. The NCDD partners have assumed responsibility for the Sustainable Digital Heritage work package. The aims of this section of the DHN's strategy plan correspond with the NCDD's mission: to ensure the long-term accessibility of digital information through the establishment of a national network of facilities.

As mentioned before, the third work programme will focus on preservation an issue, following the lines alongside the NCDD was used to work. The work programme consists of eight projects centered on three themes: (1) Scalable and usable facilities; (2) Transparent cost structure; and (3) Roles and responsibilities in collection building. A number of projects also involve the use of case studies. Each of these projects contributes to the goals of the programme, and consequently the overall mission of the Dutch Coalition on Digital Preservation. The projects will be conducted from mid-2015 to late 2016.

4. THE PROJECTS

The objective of the preservation programme is to create, through cross-domain collaboration, a shared infrastructure that guarantees sustainable access to digital information. The assumption is that this cooperation will lead to an increased effectiveness, greater efficiency and cost reductions. The programme consists of a set of eight projects and five case studies, all bringing in the bits and pieces of the jigsaw puzzle of a national distributed infrastructure. This distributed infrastructure is the focal point of the programme and all other projects add to this. To sum up some of the projects with the main results.

4.1 A network of Distributed Facilities

The Distributed facilities project builds on the results and recommendations of the Survey into a national infrastructure for sustained access to digital information [15]. Starting point are the preservation facilities already in place at the large cultural heritage institutes in The Netherlands. The project intends to create a catalogue of services, which is based on a model developed in the above-mentioned survey. In this model a distributed network of nationwide facilities is described, involving all infrastructural elements needed for preservation purposes. As we hope that these existing facilities will find used usage, they need to be catalogued and to be pointed to, so more institutions in the same sector, or by institutions in different sectors could find their way towards these facilities. However, the existing facilities are not sufficient and the project supports the establishment of new ones. These are facilities in specific areas as Architecture, digital Arts and Photography. This part of the projects represents the supply side of a national infrastructure. On the other side is the demand. Organizations of smaller scale with digital collections to be preserved. Not able to develop their own preservation systems and infrastructures. These organizations should be using the infrastructures in place. To be able to do so they need to have a wider understanding of greater needs regarding digital preservation. Within the project tools will be developed which help organizations in finding their way in the large forest of systems and services. That means checklists, guidelines and finding aids. For many organizations this will be a huge step towards maturity. Many organizations are just not aware yet at what point in their own development they are. What is their maturity level and are they capable to deal with preservation questions? In order to help and monitor the level of maturity of individual organizations a tool is developed with which institutions can evaluate themselves using the Digital sustainability score model. This model is based on a range of questions regarding issues as policy, collection development, staff knowledge, costs, and preservation levels.

This Digital sustainability score model will help organizations not only in finding out at what point in their professional development they are, but it will help them indicate the issues they need to address and the steps they need to take.

4.2 Trust

Professional development and maturity development is closely related to another topic and project in the program, that of being a trustworthiness digital repository.

Archives, museums and libraries manage a growing number of our society's digital products in their e-depots. All stakeholders must be able to place their trust in the managers of these digital collections, including those in the field of digital heritage. Managers must ensure that digital heritage collections are secured and being kept accessible for the long term. In order to provide a measure for this necessary trust, a number of certification instruments for e-depots have been developed.

Within the Certification project we will stimulate, promote and support the certification of long-term digital repositories in the Netherlands. The project will deliver a roadmap for certification of Dutch repositories. This roadmap is based on the three main instruments for certification: DSA [17], DIN [18] and ISO16363 [19]. We believe that organizations should start with the basic level, the Data Seal of Approval as a first step towards trustworthiness. Not only the usual suspects involved already in preserving digital collections should be aware of certification steps, also the smaller institutes should notice the aspects of dealing with trustworthiness solutions. So we explicitly focus our attention to the commercial players in the Dutch digital preservation field. Companies offering preservation solutions are part of the roadmap. But the large Cultural Heritage institutes as the National Library and the National Archives should lead the way.

4.3 Persistent Identifiers

A third project to highlight deal with the sustainability of the accessibility of digital information. So this project focuses on persistent identifiers. The main goals of the project are firstly to raise awareness among cultural heritage institutions on the subject of persistent identifiers, secondly to develop a business model for a persistent identifier service especially for smaller cultural heritage organizations, and lastly to set up some show cases. Within this project a strategy for communications is developed in which steps and instruments are defined to raise awareness on the topic. The project also resulted in a decision tree for cultural heritage organizations to guide them through the process of selecting a particular type of Persistent Identifier (Handle, DOI or NBN:URN). With this so called PID-helper tool cultural heritage institutes learn more on the topic and are helped with finding solutions which fit their needs [20].

Created more awareness and having a helper tool is only a first, but important, step. Next step will be the implementation of services providing cultural heritage institutes with persistent identifiers. The approach of the project is a national approach, following the strategic lines of the NCDD. So, implementing persistent identifiers is not an individual implementation on organisational level, but scalable implementation. So a vendor oriented approach is chosen. This means that we will stimulate vendors building in facilities for the different PID solutions. There are several good examples of the implementation already available on this level. The National Museum of World Cultures [21] has an agreement with the developer of The Museum System (TMS) to build in a persistent identifier solution in the collection management system they are currently using. By means of this single agreement also other TMS users are able to use this service.

Also other vendors are discussing the development of PID services in their collection- and document management systems. Within the framework of the project we are discussing this with a group of Dutch vendors. This should result in the development of persistent identifier facilities to be built in into the main systems in use in Dutch heritage organizations (archives and museums).

4.4 Costs

We want cultural heritage organizations to connect to a network of preservation services, we want them to use persistent identifiers, and we want them to become more mature regarding digital preservation. But this comes with a cost. The desirability of sustained access to digital collections is obvious. The exact costs of achieving this goal, however, are as yet unclear. This lack of insight into the costs, benefits and the business case complicates the realization of sustained access within an organization.

This project builds on the conclusions of the investigation into a national infrastructure for sustained access on the results of the European 4C project, including the roadmap for cost-effective sustained access and the Curation Cost Exchange tool [22].

The project aims to get more clarification on the costs involved in making digital collections more sustainable and provide permanent access to them. To this end, the project is working on a list of indicators, gathered using the 4C project's Cost Exchange Tool. With at least 40 institutions from various domains providing cost figures, a benchmark is being created, allowing institutions to compare their costs and expenses at different stages. In addition, the project will produce recommendations and guidelines for institutions occupying key positions within the Digital Heritage Network, supporting them in including digital preservation costs in their budgets as well as coordinating these budgets amongst each other.

5. RESULTS

These are some examples of the projects carried out with the preservation program. The program at large consists of eight projects and five case studies. These case studies feed into the main goals of the program and projects in a way that they are proof of concept cases or cases focusing on very specific topics. One of the cases is a case on Digital Archaeology, digging up the "Digitale stad" which was one of the first examples of community building on the web [23]. Within another case study a research on emulation of digital art stored on cd-roms is carried out. Within the project different emulation tools are tested on a collection of cd-roms containing works of digital art.

The presentation of a national strategy and the establishment of three Work Programs are an important development, which brings many existing initiatives and plans together. This is a start of an integrated approach for access to and preservation of Dutch digital heritage. The timing is perfect as there is a growing community of professionals involved in digital preservation. The level of knowledge exchange and the willingness to collaborate is growing too. The program on sustainable digital heritage is facilitating and stimulating knowledge exchange and collaboration by means of the development of a network of professionals. This is a network of people working in the field of digital preservation and willing to share their expertise with others. As there is a growing amount of professionals, but also many others still in need of knowledge, we have to organize this within a more formalized network. One of the instruments within this network will be a digital learning environment. This is an online training environment to be used by professionals to learn and institutes to become more mature. So they will be able for the next steps to be taken.

6. ACKNOWLEDGMENTS

Many thanks to all the project managers who are doing the actual work within their projects. Barbara Sierman, Margriet van Gorsel, Gaby Wijers, Tjarda de Haan, Joost van der Nat, Frans Neggers, Kees Waterman, Arnoud Goos, Walter Swagemakers, Gijsbert Kruithof, Patrick Aerts and Carolien van Zuilekom.

7. REFERENCES

[1] http://www.dpconline.org/about/dpc-history

[2] http://www.dnb.de/EN/Wir/Projekte/Abgeschlossen/nestor.html

[3] http://www.ncdd.nl

[4] http://www.digitalpreservation.gov/ndsa/about.html

[5] http://openpreservation.org

[6] https://www.prestocentre.org

[7] Sharon Webb, Aileen O'Carroll The process of building a national trusted digital repository: solving the federation problem. Ipres 2014 Proceedings, https://phaidra.univie.ac.at/detail_object/o:378066

[8] Darryl Mead: Shaping a national consortium for digital preservation. iPRES 2014 Proceedings, https://phaidra.univie.ac.at/detail_object/o:378066

[9] Towards Preserving Cultural Heritage of Finland Heikki Helin, Kimmo Koivunen, Juha Lehtonen, Kuisma Lehtonen, 2012 NBN: http://nbn.depositolegale.it/urn:nbn:it:frd-9299

[10] http://4cproject.eu

[11] https://fedora.phaidra.univie.ac.at/fedora/get/o:429524/bdef:Content/get, p. 159-163

[12] http://www.ncdd.nl/documents/NCDDToekomstDEF2009.pdf

[13] http://www.ncdd.nl/documents/NCDDToekomst_2_Strategischeagenda.pdf

[14] http://www.ncdd.nl/en/ncdd-projects/projects-2014-2015/

[15] http://ncdd.nl/site/wp-content/uploads/2014/06/summary_NCDD_research_DEFWEB.pdf

[16] http://www.den.nl/art/uploads/files/Publicaties/20150608_Nationale_strategie_digitaal_erfgoed_Engels.pdf

[17] http://www.datasealofapproval.org/en/

[18] http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.html

[19] http://www.iso16363.org

[20] http://www.ncdd.nl/pid/

[21] http://collectie.wereldculturen.nl/default.aspx?lang=en

[22] http://www.curationexchange.org

[23] http://hart.amsterdammuseum.nl/nl/page/521/re-dds

Status of CELLAR: Update from an IETF Working Group for Matroska and FFV1

Ashley Blewer
ashley.blewer@gmail.com

Dave Rice
dave@dericed.com

ABSTRACT

The open, independent, and international standards organization Internet Engineering Task Force (IETF) has chartered a working group. It is named "Codec Encoding for LossLess Archiving and Realtime transmission" (CELLAR) and aims to develop specifications for a lossless audiovisual file format for use in archival environments and transmission. It consists of the combination of the audiovisual container Matroska, lossless video codec FFV1, and lossless audio codec FLAC. This paper reviews the status of this on-going development and thereby provides an overview of the challenges and intricacies of audiovisual specification development.

Keywords

Format standardization; audiovisual file formats; digital preservation

1. INTRODUCTION

This paper reviews the status of the ongoing work within the Internet Engineering Task Force (IETF)'s *Codec Encoding for LossLess Archiving and Realtime transmission* working group (CELLAR). The working group is tasked with the standardization of three audiovisual formats: Matroska, FFV1, and FLAC. The authors will provide an overview of the challenges, intricacies, and progress of specification development for audiovisual formats. Topics include an overview of the benefits of open standards within the context of digital preservation, methods for advocating for and supporting implementation of standards, and the relationships between specification development and development of validation software.

2. OPEN FORMATS

Matroska, FFV1, and FLAC are open file formats. Their specifications are freely available and openly licensed, continued development is open and available to the public, historical context and conversations surrounding the specification are open access, and use of the formats or their specifications is without charge and can be used by any person. Anyone can improve upon the standards body, contingent only on the standards body to collectively approve of changes.

Matroska as an audiovisual file format has been in use since 2002, with widespread internet usage. Matroska is based upon Extensible Binary Meta Language (a binary equivalent of XML) and is the foundation of Google's webm format -- a file format optimized specifically for web-streaming. Some of Matroska's features -- such as subtitle management, chaptering, extensible structured metadata, file attachments, and broad support of audiovisual encodings -- have facilitated its adoption in a number of media communities. Matroska has also been implemented into many home media environments such as Xbox and Playstation and works "out of the box" in the Windows 10 operating system.

The Matroska wrapper is organized into top-level sectional elements for the storage of attachments, chapter information, metadata and tags, indexes, track descriptions, and encoding audiovisual data. Each element may have a dedicated checksum associated with it, which is one of the important reasons why it is deemed such a suitable format for digital preservation. With embedded checksums, a specific section of a Matroska file can be checked for errors independently, which means error detection can be more specific to the error's region (as opposed to having to identify the error within the entire file). For example, a checksum mismatch specific to the descriptive metadata section of the file can be assessed and corrected without requiring to do quality control and analysis on the file's content streams. The Matroska format features embeddable technical and descriptive metadata so that contextual information about the file can be embedded within the file itself, not just provided alongside in a different type of document.

FFV1 is an efficient, lossless video encoding that is designed in a manner responsive to the requirements of digital preservation. FFV1 has rapid traction in both the development and digital preservation communities and is widely and freely distributed with the ubiquitous ffmpeg and libav libraries for video processing. FFV1's lossless compression algorithm allows uncompressed video to be reduced in filesize without loss of quality while adding self-description, fixity, and error resilience mechanisms. FFV1 version 3 is a very flexible codec, allowing adjustments to the encoding process based on different priorities such as size efficiency, data resilience, or encoding speed. FFV1 is a strong candidate for video files undergoing file format normalization prior to the OAIS-compliant repository ingestion phase. For example Artefactual's Archivematica (a free and open-source digital preservation system) uses FFV1 and Matroska as a default normalization strategy for acquired audiovisual content and recommends pre- and post-normalization FFV1+MKV validation methods [4] [8].

FLAC is a lossless audio codec that has seen widespread adoption in a number of different applications. FLAC features embedded CRC checksums per audio frame, but also contains an md5 checksum of the audio stream should decode to. Another benefit of FLAC is that it can store non-audio chunks of data embedded in the source WAVE file, such as descriptive metadata. Since FLAC is designed to store foreign data (using the --keep-foreign-metadata option), it is feasible to encode a valid WAV file to FLAC (which adds several fixity features while reducing size) and then extract the FLAC back to recreate the original WAV file bit for bit. Tools such as the flac utility and ffmpeg can analyze a FLAC file to identify and locate any digital corruption through the use of the format's internal fixity features.

3. SPECIFICATION-BASED VALIDATION

Developers of open source software have been building tools based on the Matroska specification for many years. MKVToolNix is a suite of software tools created to work specifically with Matroska files designed by Moritz Bunkus, a core developer of Matroska and EBML. A part of mkvtoolnix is mkvalidator, which is described as “a simple command line tool to verify Matroska and WebM files for spec conformance” [3]. To facilitate that the specification is well-interpreted by the developers of tools that implement it, the mkvalidator tool provides a programmatic assessment of the validity of a Matroska implementation, whereas the specification itself is meant for a human reader. The co-development of an official specification and an official validator provides a means for both humans and computers to assess and interpret the quality of a Matroska deployment. This co-development of the specification and validation tools should be considered as a model in the specification of other file formats as well.

MediaConch is software currently being developed as part of the PREFORMA project, co-funded by the European Commission. The PREFORMA consortium describes the goal “is to give memory institutions full control of the process of the conformity tests of files to be ingested into archives” [7]. The goal of the PREFORMA project is to create open source software for the most eminent archival-grade media formats: PDF, TIFF, Matroska, FFV1 video, and LPCM audio (with MediaConch focusing on the latter three). These software packages focus on the validation and conformance checking of files against their official specifications. Investigation into the development of this software has sparked conversations on the related format list-servs (Matroska-devel, ffmpeg-devel, and libav-devel) and in other public platforms like GitHub. This investigator and conservation helped raise awareness of the state of the existing specification documents and need for more format and structure standardization processes through an established open standards organization. With a collaboration between related developer and archival user communications a proposal for a working group focused on lossless audiovisual formats was submitted for the consideration of the IETF, which would become the cellar working group.

The MediaArea team (developers of MediaConch) has been working on understanding the specific details of each segment of an archival video standard, sometimes down to the bit-level, in order to develop a comprehensive conformance checker. MediaArea has previously developed Mediainfo, a command-line software application prolifically used in media archives to quickly assess file information, and MediaTrace, developed with MoMA to provide bit-level analysis on media files.

4. EARLY STANDARDIZATION WORK

Matroska and EBML were developed from the beginning with standardization in mind. The conceptual file formats, the documentation, and associated software and libraries were developed and implemented simultaneously by the same core team. The authors of Matroska documentation were also developing validation tools such as mkvalidator, so that there was both a human-readable and programmatic methods to test if a produced Matroska file adhered to the specification or not. With other file formats, the specification and validation tools are generally developed separately by distinct teams. As lead contributors to Matroska’s core libraries and validation tools are written by the same authors of the specification, there is an opportunity for the interpretation of the specification to be very clear and precise.

Matroska’s history contained many pushes in the direction of more official standardization. In 2004 (two years after the origin of Matroska), Martin Nilsson produced an RFC draft of EBML, which extensively documented the format in Augmented Backus-Naur Form (ABNF) [6]. This draft was not published by the IETF but remained on the Matroska site as supporting documentation. Also in 2004, Dean Scarff provided draft documentation for a concept of the EBML Schema. An EBML Schema would be analogous to the XML Schema for EBML Documents and could provide a standardized structure to define EBML Document Types such as Matroska and webm. Additionally extending feature support and clarifications to documentation would be ongoing themes of the development listserv.

FFV1 was initially designed and incorporated into FFmpeg in 2003 as an experimental codec. Early documentation may be seen in the Internet Archive [5]. In 2006, FFV1 was marked as stable and gained use as a lossless intermediate codec to allow video to be processed and saved to a file without impactful encoding loss or the large sizes of uncompressed video. Between 2006 and 2010 FFV1 performed favorably in lossless video codec comparisons and found some early adoption in archives. However, at the time FFV1 had notable disadvantages compared to other lossless encodings used in preservation such as JPEG2000 and Lagarith, including a lack of support for 10 bit video, need for optimization, and crucially-needed documentation and standardization efforts.

From 2010 through 2015 FFV1 underwent significant developments and increased archival integration. Michael Niedermayer, the lead format developer, significantly expanded the documentation and released FFV1 version 3, which added embedded checksums, self-description features, improved speeds with multi-threading, error resilience features, and other features that improves the efficiency of the encoding in preservation contexts. Kieran Kuhnya, Georg Lippitsch, Luca Barbato, Vittorio Giovara, Paul Mahol, Carl Eugen Hoyos and many others contributed to the development and optimization of FFV1. In 2012, work on the specification moved to more collaborative environments in a GitHub repository. During this time, archival experimentation and implementation with FFV1 expanded and many archivists (including the authors of this paper) actively participated in supporting the testing and development of FFV1’s codec and documentations.

Michael Niedermayer began documenting a specification for the format and added several features specific to preservation usage. Version 3 is highly self-descriptive and stores its own information regarding field dominance, aspect ratio, and color space so that it is not reliant on a container format alone to store this information. Other streams that rely heavily on their container for technical description often face interoperability challenges.

Much like Matroska, despite the widespread usage, the FLAC file format had not been through a process of standardization in a standards body. However the FLAC development community has authored and maintains a comprehensive specification on the FLAC website.

5. STANDARDIZATION

The IETF

The Internet Engineering Task Force (IETF) is an open and independent international standards organization, known for the development of standards for the Internet protocol suite (TCP/IP), file transfer protocol (FTP), and protocols that

compose the Simple Mail Transfer Protocol (SMTP). IETF’s parent organization is the Internet Society (ISOC), an international, non-profit organization that has set out to “make the world a better place” by “connecting the world, working with others, and advocating for equal access to the Internet” [2]. Much of the standardization work shepherded by IETF focuses on the development of standards of and is related to the transmission of information between systems in an efficient manner without error or data loss.

The working methods of IETF promote and ensure a high degree of transparency so that anyone is able to look upon processes underway and to participate within them. Communication is organized into a system of publicly accessible mailing lists, document trackers, and chatrooms. The IETF’s conferences (held three times per year) include audiovisual streams, IRC streams, and an in-room facilitator for remote participants to efficiently invite and enable participants in the process.

PREFORMA

The PREFORMA Project is a Pre-Commercial Procurement (PCP) project started in 2014 and co-funded by the European Commission under its FP7-ICT Programme. The project responds to the challenge of implementing good quality standardised file formats within preservation environments with a particular focus on providing memory institutions with control over conformance and validation testing of those file formats. Along with PDF and TIFF, the PREFORMA administrators selected Matroska, FFV1, and LPCM as open file formats of particular interest to preservation communities and selected MediaArea to develop conformance tools for those formats.

In early planning, MediaArea’s team (including the authors of this paper) noted the particular challenges in developing conformance tools for file formats whose specifications had not yet been subject to the procedures and protocols of a standards body. PREFORMA’s network of developers and memory institutions provided an environment supportive of collaboration between developers, specification authors, and archivists. Format maintainers, developers, and archivists collaborated to participate and encourage work on Matroska and FFV1 within an open and inclusive standards organization.

The IETF as a Standards Body for Audiovisual Preservation?

Through consensus with participating communities and public discussion, the IETF was selected as the most suitable standards body with which to standardize FFV1 and Matroska due in part to its open nature, transparent standardization process, facilitation of accessibility, and organizational credibility. IETF lacks paywalls and licensing barriers for accomplished and published works. IETF provides ability for all interested persons (members or not) to participate via multiple open channels. Additionally the related developer communities of ffmpeg-devel, libav-devel, and matroska-devel were well familiar with IETF either from involvement in earlier standardization efforts and IETF’s expanding role in standardizing audiovisual formats, such as OGG, VP8, and Opus.

Participants from Matroska, FFmpeg, PREFORMA, MediaArea and many other communities collaborated to propose the formation of an IETF working group to standardize lossless audiovisual file formats for preservation. Tessa Fallon presented a draft charter at the dispatch working group meeting at IETF93. The IETF approved the charter for the working group, named CELLAR (Codec Encoding for LossLess Archiving and Realtime transmission). The opening sentences of the CELLAR charter read as follows: “The preservation of audiovisual materials faces challenges from technological obsolescence, analog media deterioration, and the use of proprietary formats that lack formal open standards. While obsolescence and material degradation are widely addressed, the standardization of open, transparent, self-descriptive, lossless formats remains an important mission to be undertaken by the open source community” [1]. CELLAR’s goal is stated as being “to develop an official internet standard for Matroska (audiovisual container), FFV1 (lossless video encoding), and FLAC (lossless audio encoding) for use in archival environments and transmission” [1]. This process involves the further testing and development of the specifications of these three formats to ensure their sturdiness, success, consensus, and maintenance long into the future.

CELLAR Happenings

The work of the CELLAR Working Group can be seen, commented upon, or contributed to in a few working spaces. The mailing list is the central location for communication and discussion on works towards the working group’s objectives. The mailing list, along with other central information pertaining to the working group, is located at: https://datatracker.ietf.org/wg/cellar/charter/
The mailing list archive is available at: https://mailarchive.ietf.org/arch/search/?email_list=cellar
At the time of publication submission on 17 July 2016, the mailing list of the working group includes the participation of 82 individuals.

For both Matroska and FFV1, the working group is building upon earlier specification work done independently by the formats’ designers and contributors. A first important step in the process was making the specifications more accessible by improving their online presence. Both Matroska and FFmpeg managed in-development specification drafts on their websites with contributions from the community. Within the IETF working group this development continues with improvements to the specifications themselves and improvements to the websites that support those specifications with the goal of allowing more collaborative work by an expanded population of developers and archivists. The FFmpeg specification webpage was formerly built in LyX. In Summer 2015, the specification was migrated to Markdown, a syntax easier to read and easily hosted on collaborative version control platform, Github. Similarly, the Matroska specification was hosted in the main Matroska website, built in Drupal. It has also been migrated to Markdown and Github to promote collaboration of specification refinement work done primarily in conversation via the CELLAR listserv.

Accomplishments via CELLAR

CELLAR work has resulted in producing valid RFCs for EBML, Matroska, and FFV1 for official consideration at IETF’s July 2016 conference. These RFCs are early draft specifications constructed through restructuring, clarifying, and building upon the existing specification as well as adding sections mandated by RFC guidelines such as security considerations, abstracts, and references.

Overall the work of cellar has fallen into three categories. 1) Meeting IETF’s documentation requirements through adding mandated sections such as security considerations, valid references, abstracts, and notations. 2) Improving existing documentation, such as rewriting and refining what has already been put into practice but needs fine-tuning. 3) Extending features to accommodate new use cases and respond to past lessors learned.

New features have been proposed and added to the updated specification, including a proposal for color management (via Google in relation to WebM), disambiguation and refined specification for timecode, and improvements to interlacement status.

Existing features require further clarification, and much work has been done in this area. This involves gather use cases, reviewing the existing specification, and fixing discrepancies between elements and clarifying the language when vague or able to be interpreted (or have been interpreted) in different ways.

Within the working group the sections of Matroska's specification that pertained to its underlying EBML format where consolidated into a EBML specification, so that the Matroska specification may build upon the EBML specification rather than act redundantly to it. The updated EBML specification includes documentation on how to define an EBML Schema which is a set of Elements with their definitions and structural requirements rendered in XML form. Matroska's documentation now defines Matroska through an EBML Schema as a type of EBML expression.

RFC drafts have been submitted in anticipation of IETF96 and the CELLAR working group meeting (held on 19 July 2016). During this meeting, the specification will be reviewed. Comments will then be discussed and implemented into the next version of the EBML RFC. There is still a long way to go in refining these RFC documents to IETF standards and consensus as can be seen in the comprehensive reviews arriving at the cellar listserv prior to the working group meeting.

The work of the cellar working group is ongoing and active. The working group provides a unique environment where archivists are working alongside developers and specification authors.

6. FORMAT RECOMMENDATIONS

The specifications of Matroska and FFV1 permit a range of flexible usage to accommodate distinct use cases and priorities. Specific uses certainly benefit from specification optimization and recommended practice. Best practices for the usage of both Matroska and FFV1 are evolving due to the work of the CELLAR working group. However the authors of this paper would like to present recommendations for optimization for current use and look to what may be useful in future refinements of FFV1 and Matroska intended specifically for digital preservation.

The benefits and security of whole-file checksums do not scale fairly for larger audiovisual files. Whereas an electronic records collection may store thousands of files in the space of a terabyte and thus manage thousands of corresponding checksums to authenticate the storage, an audiovisual collection may use a terabyte to occupy a few dozen files. The larger the file is, the less effective a checksum mismatch is at clarifying the extent and location of the error. Both FFV1 and Matroska incorporate fixity features so that pieces of the data utilize their own checksums.

Matroska adopts of feature of its foundational format EBML, which supports nested checksum elements into any structural element container. The EBML specification states "All Top-Level Elements of an EBML Document SHOULD include a CRC-32 Element as a Child Element." This enables attachments, track metadata, description metadata, audiovisual data and all other sections to have the ability to manage their own checksum. This allows a much more granular and targeted use of checksums and also enables parts of the file to be changed while maintaining the fixity of the other parts. For instance a Matroska file may store audiovisual content, attached images of the source video tape, and logs of the creation of the file. Add a later stage in the archival life of the Matroska file, a quality control report may be created about the file and then itself stored within the file without affected the fixity of the audiovisual data.

FFV1 version 3 mandates the storage of checksums within each frame so that the decoder may know precisely if a frame is valid or invalid. Optionally FFV1 version 3 can incorporate checksums into each slices of the frame. In this case, if the data is corrupted the decoder can know what region of the frame is damaged and conceal it by duplicating pixels from the previous valid frame into the corrupted space. FFV1 is able to re-use contextual information from frame to frame as a way of reducing its data rate; however the re-use of context across frames can reduce the error resilience of FFV1. In preservation it is recommended that all FFV1 frames are encoded as self-dependent so that they are not dependent on information from another field. This is done by setting the GOP (group of pictures) size of the FFV1 encoding to 1.

FFV1 encodings are generally much faster than other lossless encodings partly because of the support of multithreaded encoding. With multithreaded encoding the frame is sliced into many slices that are encoded through separate processes and merged back into a frame. Encoding with slices also reduces the visual effects of data corruption by regionalizing damage to a smaller contained area. It is recommended to use a higher slice count such as 24 or 30 while encoding to benefit from these features.

FFV1 version 3 incorporates significant preservation features over the prior versions. Within version 3, the micro versions of 3.1, 3.2, and 3.3 were experimental and version 3.4 was the first stable release. So specifically, version 3.4 is recommended.

Both FFV1 and Matroska incorporate a significant amount of self-description. It is recommended that such metadata be declared specifically (rather than noting an 'undetermined' value) and that the metadata is consistent between the Matroska container and FFV1 encoding. For instance FFV1's picture_structure value should clarify the interlacement and not be set to 'undetermined' unless it truly is undetermined. Additionally FFV1's sar_num and sar_den (which document sample aspect ratio) should be explicitly set rather than set as '0' which would indicate an unknown sample aspect ratio.

As videotapes are digitized there is a lot of contextual information to clarify. Videotape players generally do not communicate values such as audio channel arrangement or aspect ratio (especially true with analog media). A videotape may have traditional default considerations, such as considering the first audio channel as left, second as right, and aspect ratio as 4/3; however, this should be clarified in the digitization process and not left to assumption. It is recommended that values such as aspect ratio and audio channel arrangement be set explicitly where possible.

Often a physical audiovisual carrier is not able to communicate the aperature or boundary of the image during digitization. For instance a 720x486 encoding of video may only contain a 704x480 active picture bordered by rows and columns of black pixels. Alternatively a film scan may include film perforations, sound track data, or the border between frames. The framing of the presentation can be clarified using Matroska's PixelCrop elements. This allows the active picture to be set according to coordinates while preservation the entirety of the encoding image. This feature can also allow black pixels from letterboxing or pillarboxing to be hidden or possibly to hide head switching or unintended video underscan from the presentation while preserving it.

Legacy videotape does not contain a method for a machine to understand where the content starts and ends. Additionally legacy videotape often contains supporting content for technical and historical reasons. For instance a 30 minute program on videotape may be included with several other minutes of color bars, informational slates, countdown, black frames, and other material not intended to be part of the presentation.

Matroska's chaptering support includes a feature called Ordered Chapters. With Ordered Chapters a user be document various intended presentations of the video. For instance, one Matroska file may contain a set of chapters that presents the entirety of a digitized videotape (including color bars, static, black frames and whatever else is present). The same file may contain another edition of chapters that presents only the featured content of the tape and skips over the colorbars and other technical video content. Players such as VLC provide means to switch between chapter-managed presentations. It is recommended to showcase the intended presentation with the default edition of chapters and provide access to the full encoding of the videotape's content via an alternate edition of chapters.

Matroska has a strong focus on managing language for subtitles, audio, and metadata. While Matroska defaults to English, it is recommended to clarify language properly, so that if a file contains many alternate audio encodings or sets of metadata that their language is properly marked.

Recommendation Summary (ffmpeg options are in backticks):

When storing content in Matroska for preservation use CRC-32 Elements in all Top-Level Elements as suggested by the EBML specification.

When encoding FFV1 for preservation include the options: `-level 3` and `-slicecrc 1` to request FFV1 version 3 with slice crcs enabled.

Use an FFV1 GOP size of 1 with `-g 1`.

Use a high slice count (at least 24) during FFV1 encoding, `-slices 24`.

Avoid setting FFV1 values of picture_structure, sar_num, sar_den to an 'unknown' value.

Use of FFV1 of at least version 3.4 (major version 3 and micro version 4).

Be as explicate and accurate as possible when storing information about aspect ratio, audio channel arrangement, presentation timeline, and language.

Consider using Order Chapters to distinguish the intended presentation of a digitized videotape from other technical content (such as color bars and countdown).

Also of these recommendations are feasible with mkclean, mkvpropedit, and ffmpeg or avconv.

7. CONCLUSIONS

Historically, many digital audiovisual formats put forth as archival standards have been proprietary and have fallen out of common usage as they become outdated, unpopular, or as the governing company loses interest in seeing the project continue. The specifications of both FFV1 and Matroska have been actively developed in an open source and open license environment that welcomes participation and review. Many of the prominent contributors and authors of these specifications also concurrently contribute to the development of open source tools to utilize, assess, or integrate these formats. As a result, the specification development isn't wholly idealistic but the design effort is tied to ongoing contributions to the main open source projects that support the specifications. The work in CELLAR to improve the specifications is an effort that parallels efforts in VLC, Libav, FFmpeg, MKVToolNix, and other open source toolsets that deploy the new aspects of the specification.

FFV1 has been at a tipping point in adoption within the preservation community. Archivematica has adopted FFV1 for lossless video normalization for long term preservation. More digitization vendors have added support for the format as well. Matroska has been under a slower adoption by archives but its features for sectional fixity, hierarchical metadata, attachments, and preservation data make it worthy for consideration. Additionally as the specification is open source and its refinement is in an active IETF working group that specifically focuses on archival use, archivists are encouraged to review and participate in this effort.

8. REFERENCES

[1] IETF Cellar Working Group, "CELLAR Charter," [Online] Available: https://datatracker.ietf.org/wg/cellar/charter/

[2] Internet Society, "Internet Society Misison," [Online] Available: http://www.internetsociety.org/who-we-are/mission/

[3] Lhomme, Steve and Bunkus, M., "mkvalidator tool," [Online]. Available: https://matroska.org/downloads/mkvalidator.html

[4] McLellan, Evelyn, "Format Policies," [Online] Available: https://wiki.archivematica.org/Format_policies

[5] Niedermayer, Michael, "Description of the FFV1 Codec," [Online] Available: https://web.archive.org/web/20030807095617/http://mplayerhq.hu/~michael/ffv1.html

[6] Nillson, Martin, "EBML RFC (Draft)," [Online] Available: https://matroska.org/technical/specs/rfc/index.html

[7] PREFORMA Consortium, "PREFORMA Future Memory Standards," [Online] Available: http://preforma-project.eu/

[8] Romkey, Sarah, "Requirements/MediaConch integration," [Online] Available: https://wiki.archivematica.org/Requirements/MediaConch_integration

9. CREDITS

Many thanks to Natalie Cadranel, Sebastian G., Kieran O'Leary, and Erwin Verbruggen for their contributions and suggestions to this article.

PREMIS 3.0 Ontology: Improving Semantic Interoperability of Preservation Metadata

Angela Di Iorio
DIAG - Department of Computer,
Control, and Management
Engineering Antonio Ruberti,
Sapienza University of Rome
Via Ariosto 25 00185, Rome, Italy
angela.diiorio@uniroma1.it

Bertrand Caron
Department of Metadata
Bibliothèque nationale de France
Quai François Mauriac
75706 Paris Cedex 13
bertrand.caron@bnf.fr

ABSTRACT

The PREMIS 3.0 Ontology Working Group is a community interested in using Semantic Web Technology to leverage systems managing the long-term preservation of digital archives.

The version 3 of the PREMIS Data Dictionary has stimulated the community to revise the current PREMIS OWL Ontology. The revision process aims not only to integrate the conceptual model with the changes defined by the new data model of the PREMIS version 3.0, but also to ease the implementation of Semantic Web Technology in the digital preservation community.

Keywords

semantic web technologies; preservation metadata; PREMIS ontology.

1. INTRODUCTION

In this article, the development work for reviewing the PREMIS OWL Ontology [4] is introduced. The PREMIS 3.0 Ontology Working Group is a community interested in using Semantic Web Technology to leverage systems managing the long-term preservation of digital archives.

The current PREMIS OWL is a semantic formalisation of the PREMIS 2.2 Data Dictionary [6] and defines a conceptual model for the metadata that a digital archive needs to know for preserving objects. In June 2015 version 3 of the PREMIS Data Dictionary [7] was released. This in turn has led to a community review of the PREMIS OWL. The review process aims not only to integrate the conceptual model with the changes defined by the data model of the PREMIS version 3.0, but also to ease the implementation of Semantic Web Technology in the digital preservation community.

The PREMIS version 3.0 changed the PREMIS Data Model and refined the description of the digital objects' Environment, a specific type of Intellectual Entity. These changes have implied the revision of the previously published ontology. The revision working group felt that a deeper revision of the existing ontology should be made. Indeed, the previous modelling work had taken as a starting point the PREMIS XML Schema, and automatically transformed it in an OWL file. The obtained ontology was thereby quite close to the Data Dictionary structure and vocabulary, though some simplifications were made to make it more RDF-friendly.

In order to go further in that direction, the PREMIS 3.0 Ontology Working Group decided to look at semantic units of the PREMIS Data Dictionary, not directly as classes and properties, but as description elements of real-world objects. In other words, the dictionary has to be turned into a formalisation of the digital preservation knowledge domain. This perspective implies some significant changes in the ontology. Nevertheless, the revision working group is performing a reconciliation between these necessary changes and the coherence with the PREMIS Data Dictionary.

2. THE EVOLUTION OF THE PREMIS PRESERVATION METADATA

The PREMIS Data Dictionary (PREMIS-DD) is built on the Open Archival Information System (OAIS) reference model (ISO 14721) [2]. The PREMIS-DD defines specifications about which metadata is necessary to preservation practices and provides directions for implementations.

The PREMIS XML schema[1] is usually provided in parallel with the PREMIS-DD for supporting the XML implementation of the preservation metadata management.

The PREMIS Data Model underlying the PREMIS-DD consists of five main information entities [6] deemed important for digital preservation purposes:

1) *Intellectual Entity*, an intellectual unit for the management and the description of the content.
2) *Object*, a discrete unit of information subject to digital preservation. The Object has three subtypes:
    a. *File* is a named and ordered sequence of bytes that is known by an operating system.
    b. *Bitstream* is contiguous or non-contiguous data within a file that has meaningful common properties for preservation purposes.
    c. *Representation* is the set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity.
3) *Event,* an action that has an impact on an Object or an Agent.
4) *Agent,* a person, organization, hardware or software associated with Events in the life of an Object, or with Rights attached to an Object.
5) *Rights,* a description of one or more rights, permissions of an Object or an Agent.

The PREMIS-DD version 3.0 was published in June 2015. The major changes and additions provided by this last version describe dependency relationships between Objects and their Environments: hardware and software needed to use digital objects.

This evolution has required two main repositions:

1) the *Intellectual Entity* is defined as a category of Object to enable additional description and linking to related PREMIS entities;
2) the *Environments* (i.e. hardware and software needed to use digital objects) are described as generic Intellectual Entities so that they can be described and preserved reusing the Object entity, as Representation, File or Bitstream.

This change allows for Environment descriptions or even their Representations to be shared. By expanding the scope beyond repository boundaries, the data interoperability among repository systems is improved, because the Environments descriptions is more granular and consistent with their original technological nature.

3. USE CASES AND SCOPE OF THE PREMIS 3.0 ONTOLOGY

The PREMIS 3.0 Ontology Working Group (WG) has initially collected use cases, that would benefit from integrating the use of PREMIS Ontology, in current RDF implementations of digital archives.

The WG has solicited the new version of PREMIS Ontology as a conceptual model for producing RDF datasets expressed in PREMIS 3.0 terms (classes and properties), to be combined with other terms defined by third-party ontologies provided in RDF Schema[2] or OWL [5]. For example, the integration with the Europeana Data Model (EDM)[3], as well as interest in using PREMIS 3.0 Ontology in systems Hydra/Fedora 4 based, by integrating it in the Portland Common Data Model (PCDM)[4], has been discussed by the group and has been considered a feasible test bed for releasing the new ontology.

The general assumption of the WG was that the objective of adopting as much as possible an approach oriented toward the interoperability with other well established ontologies would generally contribute to increase the interoperability of digital archives aiming to use the PREMIS 3.0 Ontology.

Other ontologies that will be considered by the WG for the integration are: PROV-O [12] for provenance information, FOAF[5] for human agents, DOAP[6] for software agents, Dublin Core for descriptive metadata, and OAI-ORE for structural relationships.

Over and above this specific goal (aiming to improve the metadata interoperability of digital repositories) a specific need for improving the interoperability of the management of preservation metadata has also arisen from the WG.

Current practices in searching resources with specific characteristics, usually rely on domain knowledge, expertise, and professional networks of categories, involved in the digital preservation. The cross-repository search can also be complicated by the interoperability problems due to different underlying data models of digital repositories. The need for developing a model connecting different RDF datasets, related to the preservation metadata domain, has led the WG to revise and integrate the current PREMIS OWL ontology.

The integration of third-party ontologies will help to overcome these limitations and to engage user communities to deeply use preservation metadata for supporting their research and to help stakeholders in improving the management of preservation metadata.

The scope of the PREMIS 3.0 Ontology is indeed to support the implementation of different Semantic Web Technology through the digital preservation community, and will support these technologies to answer questions that could arise out of the community (users and stakeholders).

The repositories using the PREMIS 3.0 Ontology as conceptual model for RDF datasets, should have the ability of answering questions like: What is the relationship between an Object and another? How many Files is a Representation made of? What is the average number of Files of all Representations? When was a Representation created? How many Representations were ingested after certain date? Which Files are JPEG images? Which Representations contain Files in PDF format?

4. THE PREVIOUS PREMIS ONTOLOGY

Starting from version 2.2, a PREMIS OWL ontology has been made available alongside the PREMIS XML Schema.

The PREMIS OWL ontology is a semantic formalisation of the PREMIS 2.2 data dictionary [6] and defines a conceptual model for the preservation information of a digital archive. The PREMIS OWL ontology has allowed the interested community to express preservation metadata in RDF, by using the conceptual model of the PREMIS-DD, and as such, it can be used to disseminate the preservation information as Linked (Open) Data [1].

The design of the PREMIS OWL [2] has tried to be coherent as much as possible to the PREMIS-DD, aiming at preserving the knowledge model. As such, the structure of the PREMIS-DD semantic units, defined by experts in the domain of the long-term digital preservation, and its translation in the XML schema, have been replicated in the PREMIS OWL.

The PREMIS OWL has addressed the problem of interoperability, deriving from the preservation policies and processes that each digital preservation archive adopts, by using the formalism of the Web Ontology Language (OWL 1) [5] [8]. In addition, 24 preservation vocabularies have been integrated, that are exposed by the Library of Congress Linked Data Service[7] and are provided as SKOS [8][9] preservation vocabularies.

The PREMIS OWL does not replace but rather complements XML in areas where RDF may be better suited, such as querying or publishing preservation metadata, or connecting repository-specific data to externally maintained registries.

At the time of its design, the PREMIS OWL has deviated from the PREMIS 2.1 Data Dictionary trying to reconcile the model differences between the XML schema and the OWL ontology[8] [2].

The principles and design deviations, as well as the OWL implementation choices have been reviewed by the PREMIS 3.0 Ontology Working Group as a starting point for modelling the PREMIS 3.0 Ontology.

[1] PREMIS Preservation Metadata XML Schema VERSION 3.0, http://www.loc.gov/standards/premis/premis.xsd

[2] RDF Schema 1.1, https://www.w3.org/TR/rdf-schema/

[3] Europeana Data Model Documentation, http://pro.europeana.eu/page/edm-documentation

[4] Hydra and the Portland Common Data Model (PCDM), https://wiki.duraspace.org/pages/viewpage.action?pageId=69011689

[5] Friend of a Friend (FOAF), http://semanticweb.org/wiki/FOAF

[6] Description of a Project (DOAP), https://github.com/edumbill/doap/wiki

[7] Library of Congress LD Service, http://id.loc.gov/vocabulary/preservation.html

[8] Public workspace for PREMIS OWL ontology, http://premisontologypublic.pbworks.com

5. PREMIS 3.0 ONTOLOGY: WORK IN PROGRESS

In order to make the new version of the ontology more compatible with Linked Data Best Practices[9], the WG followed a similar approach to the one adopted for the revision[10] of the Bibframe[11] ontology. The following principles were agreed upon, though on specific points the working group may decide against them. Some of them were already followed in the previous version of the ontology, some others were not and their adoption may bring important changes in the next version.

5.1 Make it Simple

Simplicity is the key to massive adoption; that is why the working group has the objective of making the ontology as simple as possible; but not simpler. Some of the following principles derive from this generic one, which should be kept in mind at any step of the modeling process.

5.2 Use PREMIS-DD as a Knowledge Base

Having a Data model is a real asset when trying to build an ontology: theoretically, it would provide classes and the Data Dictionary properties. In the case of PREMIS, RDF modeling has to consider other concepts which are in the preservation domain (generally existing as semantic containers in the Data Dictionary) but do not appear in the Data Model, e.g., Signature, Outcome, Registry, etc. Thus the ontology cannot be an exact transcription of the PREMIS Data Dictionary in OWL. The WG had to reconcile two opposite directions: sticking to the PREMIS Data Dictionary or introducing conceptual discrepancies with it in order to reflect more faithfully the preservation activities and to respect ontology design principles.

The PREMIS Data Dictionary is built on the principle of technical neutrality. It gives a list of pieces of information to record without any constraint on where and how to record it. According to the PREMIS conformance principles, implementers can store information anywhere, with any structure and any element names, provided that they can establish an exact mapping between their data and PREMIS semantic units. That is why the WG considers scope, concepts, and intent provided by the Data Dictionary, but feels free to differ regarding the names and structure of the ontology.

As said above, the Data Dictionary provides pieces of information, whereas the ontology describes real-world objects and organizes knowledge on these objects. One example is about semantic containers, a mechanism extensively used by PREMIS to group together related pieces of information. Systematically transcribing them into the ontology would create extra levels of indirection and make data processing more difficult. If high-level containers become classes (e.g. the fixity semantic container becomes a `premis:Fixity` class, as the "fixity" is a real-world concept), for semantic containers of lower level (e.g., formatDesignation, which is only used to group the format name and its version). Their existence as classes in the next version of the ontology is still being debated.

5.3 Re-use Pieces of Existing Ontologies

The scope of the ontology – preservation – covers many other domains: technical characteristics of files, software description, cryptographic functions, structural relationships between digital resources, digital signature, etc. Many of these domains have already defined ontologies whose re-use is worth investigating. Re-using existing vocabularies is one of the most important notions of the semantic web and is agreed best practice, as it is saving time not as much for ontology designers but mainly for developers and consumers. Instead of distrusting other ontologies because of their potential evolution, relying on the expertise of their maintainers seems a better option.

This principle is probably the main difference between the new approach and the previous published ontology, in which re-using vocabularies had been avoided to stick to the Data Dictionary semantic units. The following elements are taking into account when examining the relevance of existing vocabularies, which is made case-by-case:

The classes of an ontology should correspond to concepts within that particular knowledge domain – if PREMIS needs elements that are not specific to the preservation domain, it should ideally pick existing elements in another domain model.

In the case of multiple possible existing ontologies, preference should be given to stable, better-known and more frequently used ones.

When considering re-using an external element, its definition must be taken into account, but also its properties, and especially domain and range, as inference processes will deduce the type of the subject and object of a re-used property. Re-use existing ontologies can thus bring more work to the ontologist but it naturally improves interoperability.

5.4 Re-use of LOC-LDS Preservation Vocabularies

Updates to existing preservation vocabularies and integrations of new ones have been performed[12] coherently with the version 3 of the PREMIS-DD and before of the WG ontology revision. Except for the vocabulary related to the Event types which is still under revision gathering the community feedback, 26 vocabularies have been released. For example, an "Environment function type" vocabulary[13] was created to provide URIs and definitions for the most common types of Environments considered by the PREMIS Editorial Committee: hardware peripheral, plugin, chip, operating system, etc.

Some of the preservation vocabularies were included in the previous version of the ontology; for example, Agent roles[14] were declared subproperties of the `premis:hasEventRelatedAgent`. The same solution was foreseen for the new version of the ontology, in order to manage two different update frequencies, as the ontology should be rather stable compared to vocabularies like software types, which are likely to be submitted to frequent changes. Nevertheless, a discrepancy appears between the ontology, whose classes and properties are designating real-world objects, and preservation vocabularies, which are authoritative vocabularies and designate a concept - they are declared as subclasses of `skos:Concept`. Importing preservation vocabularies which are a collection of simple thesauri and use such terms as subclasses of real-world objects, or re-declaring in the ontology classes and properties as real-world objects designated by these terms, is still a pending question.

5.5 Establish Equivalent Terms

When re-using is not possible, another way of improving vocabularies interoperability is to declare equivalent terms. In the case the direct re-use of an external element is not chosen because of the element being broader or not directly equivalent, linking the PREMIS element to the external one can be done with properties like (from the meaningful to the most lightweight) the OWL equivalentClass or the RDFS subClassOf and seeAlso properties. For example, the PREMIS class for software Environments could be declared a subclass of the DOAP Project class, so that consumers aware of the DOAP ontology can deduce information about PREMIS software Environments.

Using these properties to link PREMIS ontology elements with elements from other existing ontologies was planned in the previous version of the ontology, though it had not been done.

5.6 Use URIs to Identify Things

Identifying a resource on the web is typically done with URIs, as strings do not provide the same assurance about uniqueness. Literals are dead-ends in linked data, as no assertions can be made from them. Consequently, instead of having a list of values to identify the type of any described entity, a best practice is to create URIs for each item inside the list. To achieve this goal, LOC-LDS preservation vocabularies are considered the reference point, because they provide URIs for terms that are commonly needed by implementers and endorsed by the PREMIS Editorial Committee.

The enumeration is not meant to be comprehensive but extensible: if the list is insufficient to some implementers, they can just coin their own URIs, more tailored to their needs, and declare them members of the corresponding list.

5.7 Follow Best Practices Naming

The names of the classes and predicates should follow best practice naming conventions. Element names should be in "CamelCase". Classes should be initial upper case noun phrases (ClassOfThing), predicates should be initial lowercase verb phrases (hasSomeRelationship). Ambiguous names should be avoided: "isPartOf" / "hasPart" is preferable to "part" which does not indicate at first sight which is the part and which is the whole. Final names of the classes and properties to be created in the ontology can be deferred until the end of the process.

This principle has been followed in the previous ontology. Nevertheless, LOC-LDS preservation vocabularies were designed to be used in different technical contexts (XML files, databases, etc.) and thus do not follow this practice (for example, the URI `http://id.loc.gov/vocabulary/preservation/environmentFunctionType/haa`, possibly abbreviated as `envFuncType:haa`, does not satisfy the requirements for the clarity mentioned above).

5.8 Provide Documentation and Guidelines

As the ontology vocabulary can differ on some points with the Data Dictionary semantic units, documenting all ontology elements and providing guidelines for expressing XML structures as RDF triples is absolutely necessary. The maintenance of documentation and guidelines should not be underestimated either.

6. APPROACH AND TOPICS UNDER DISCUSSION

The PREMIS 3.0 Ontology Working Group has selected specific topics on which focusing the revision process and discussing the design of the PREMIS 3.0 Ontology.

In line with the principles adopted, the general approach has been to revise the conceptual connection between the ontology and the concepts expressed by related LOC-LDS controlled vocabularies (see Section 5.4).

Furthermore, some topics have catalysed questions about choices to be made in developing the conceptual model of the Ontology. Below is provided a list of questions arising around topics and that are being discussed by the WG:

Identifiers and URIs: what is the difference between URIs for identifying RDF resources with respect to the Identifiers semantic unit widely used in the PREMIS-DD? And what is the Identifier entity? Do we need an Identifier class given that identifiers in RDF are the URIs that unambiguously designate resources?

Preservation Level: how is preservation level decided? Are there other entities not included in the PREMIS-DD that could help us modelling PreservationLevel, like for example a top-level Preservation Policy class? Are both preservation levels types and preservation level roles subclasses of Policy? Would it be useful to link them to a policy assignment Event to keep track of their change through migrations?

Significant Properties: are significant properties actually globally true features of the object, or are they assigned by different preservation policies? Would it be useful to link them to a policy assignment Event to keep track of their change through migrations?

The values of significant properties appear to be free text. Is this even useful to record, when it is not machine actionable? Could it just be a `premis:note`?

Environment: has the Objects' environment to be re-modelled, based on the changes in the PREMIS-DD version 3.0?

Agent: Is it possible to define the equivalence of the Agent class with the Agent class defined by the PROV-O or FOAF?

7. FUTURE DEVELOPMENTS

The answers to the listed questions and the choices made for the design of the ontology will lead to the release of the PREMIS 3.0 Ontology. The publication of the new version of the ontology will take into account the provision of proper documentation, following the principles established by the WG.

In addition, the engagement of a wider community, by providing different serialization formats for allowing a wider re-use in the semantic web community will be also considered: the OWL 2 [11] RDF/XML serialization will be released for being used by conforming OWL 2 tools[15]. Additional formats, more readable by the implementers like the Turtle[16] or OWL 2 Functional syntax[17] will be also provided.

8. ACKNOWLEDGMENTS

We would like to thank all members of the PREMIS 3.0 Ontology WG that participated to the group activities and contributed at different level to this article. We thank Rebecca Guenther, Robert Sanderson, Esmé Cowles, Charles Blair, Karen

[9] Best Practices for Publishing Linked Data, http://www.w3.org/TR/ld-bp/

[10] Rob Sanderson, Bibframe Analysis, bit.ly/bibframe-analysis

[11] Bibliographic Framework Initiative, https://www.loc.gov/bibframe/

[12] Revised and new preservation vocabularies, http://id.loc.gov/vocabulary/preservation.html

[13] LOC-CDS Environment function type, http://id.loc.gov/vocabulary/preservation/environmentFunctionType

[14] LOC-CDS Agent role in relation to and Event, http://id.loc.gov/vocabulary/preservation/eventRelatedAgentRole

[15] OWL 2 serialization technical requirements, https://www.w3.org/TR/owl2-overview#Syntaxes

[16] RDF 1.1 Turtle - Terse RDF Triple Language, https://www.w3.org/TR/turtle/

[17] OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition), https://www.w3.org/TR/owl2-syntax/

Estlund, Gloria Gonzalez, Arwen Hutt, Evelyn McLellan, Elizabeth Roke, Ayla Stein, Peter McKinney, Ben Fino-Radin.

REFERENCES

[1] T. Berners-Lee. 2006. Linked data-design issues.

[2] Consultative Committee for Space Data. 2012. Reference Model for an Open Archival Information System (OAIS), Recommended Practice CCSDS 650.0-M-2 Magenta Book.

[3] S. Coppens, S. Peyrard, R. Guenther, K. Ford, T. Creighton. 2011. PREMIS OWL: Introduction, Implementation Guidelines & Best Practices.

[4] S. Coppens, R. Verborgh, S. Peyrard, K. Ford, T. Creighton, R. Guenther, E. Mannens, R. Walle. 2015. Premis owl. Int. J. Digit. Libr., 15(2-4):87-101.

[5] D. L. McGuinness, F. Van Harmelen, et al. Owl web ontology language overview. 2004. W3C recommendation, 10(10):2004.

[6] PREMIS Editorial Committee. 2012. PREMIS Data Dictionary for Preservation Metadata version 2.2.

[7] PREMIS Editorial Committee. 2015. PREMIS Data Dictionary for Preservation Metadata version 3.0.

[8] W3C. 2004. OWL Web Ontology Language Overview.

[9] W3C. 2009. SKOS Simple Knowledge Organization System Primer.

[10] W3C. 2009. SKOS Simple Knowledge Organization System Reference.

[11] W3C. 2012. OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition).

[12] W3C. 2013. PROV-O: The PROV Ontology.

Exhibiting Digital Art via Emulation

Boot-to-Emulator with the EMiL Kiosk System

Dragan Espenschied
Rhizome
235 Bowery
New York, NY, U.S.
dragan.espenschied@rhizome.org

Oleg Stobbe, Thomas Liebetraut and Klaus Rechert
University of Freiburg
Hermann-Herder Str. 10
79104 Freiburg, Germany
{first.lastname}@rz.uni-freiburg.de

ABSTRACT

The availability and accessibility of digital artworks is closely tied to a technical platform, which becomes quickly unavailable due to a fast technical life-cycle. One approach to keep digital artworks performing is to replace physical hardware parts with emulation. Preparing an emulator to publicly display digital art is typically time-consuming and, more importantly, usually done on a case-by-case basis, making each installation a unique and costly effort.

We present an adaptation of the Emulation as a Service framework to be deployed on a self-contained USB-stick, booting directly into a prepared emulated environment. Furthermore, we report from practical experiences using the system in two museum exhibitions.

Keywords

Emulation, Digital Art, Net Art, public access, exhibition

1. INTRODUCTION

With a growing amount of digital artworks relying on outdated hardware to perform, art museums and galleries are struggling to publicly present them for example in historical exhibitions dedicated to a certain period or movement within digital art.

Especially net art, with its requirements for active usage and network connections, has been posing challenges since its inception in the 1990's, with works being difficult to exhibit long before they became historical. While many art institutions have consequently moved digital art into the space of special events, those committed to presenting historical perspectives of digital art have created a demand for sophisticated emulation setups. Preparing an emulator to publicly display digital art is typically time-consuming and, more importantly, usually done on a case-by-case basis, making each installation a unique effort.

When artistic or curatorial intents demand that not only the artworks' computational part is retained, but also the "look & feel" of certain pieces of historical hardware is required (monitors, input devices), exhibitions can grow into hard to manage, very fragile undertakings, placing an undesirable strain on institutions regarding technical and personal resources. The main reason for this is not so much the required hardware, which in the case of net art has nothing unique to it and is easily replaced (e.g. no monitors have to be physically manipulated). Problems of scale rather arise on the computing side, when multiple, technically differing computer systems have to be configured at the software level to behave in the intended way while being replaceable in the case of hardware failure.

To address the aforementioned challenges, this paper presents a technical solution and a workflow based on the Emulation as a Service framework, making use of a range of emulators and guided web-workflows for preparing a specific emulation setup. Furthermore, we present an adaptation of the EaaS framework to be deployed on an self-contained USB-stick, booting directly into a prepared emulated environment. Finally, we report from practical experiences using the system in two museum exhibitions.

2. EXHIBITING NET ART

Net art is an art form with its root in the early 1990's, using mainly the World Wide Web as an artistic medium. Most net art has not been created to be presented in a bricks-and-mortar gallery, but with the Web itself being the point of contact with the audience. As that, net art is one of the least self-contained digital art forms, with lots of complex technical and infrastructural dependencies required for its performance.

Yet different cultural institutions have brought net art into their galleries, due to its cultural significance. Additionally, institutions have the chance to present historical net art that has become inaccessible or distorted on the Web, because of data loss or changes in consumer devices and software since a work was created. Gallery visitors can be presented historically accurate settings, adding legacy software and hardware, something that the Web can not offer.

Three main ways to publicly present net art in physical space have been established very early in the practice's history:

- *The unmodified, connected computer*
  Example: Documenta X, Kassel, 1997
  Off-the-shelf consumer devices, with typical software required to access the web, are used to present artworks. Visitors of the space see the familiar devices and interfaces and are able to fully interact with them. This matches the intended environment for the artworks, but inevitably leads to the audience modifying the setups to the point when they become technically un-usable or don't behave as intended in a very short amount of time. Gallery goers reading their email on gallery computers instead of focusing on the art has been a common sight, and still represents the least

problematic ”unintended use”.

- *Locked-down Kiosk systems*
Example: net_condition, ZKM, Karlsruhe, 1999
To prevent the aforementioned ”unintended use”, kiosk systems with very restricted interfaces are used so the audience has no way of modifying the computer set-ups or even ”surfing away” from the artwork. While this makes the exhibition practical, in many cases these restrictions hamper the affect of the artwork, for example by removing visible URLs or common elements the works refer to, like navigation buttons or widgets of the operating system. Given that most net artworks are not created for kiosk systems, there is also no guarantee that they would even perform as intended.

- *Video documentation*
Example: most art fairs ever since
In the face of the aforementioned complexity, many institutions fall back on digital video to show any kind of digital art. While this is definitely the easiest approach, for various reasons it is in many cases unable to transport or represent an artwork that was not created as a video.

While institutions have to weigh the pros and cons for each of these presentation forms, legacy artworks add yet another dimension of issues: If an artwork benefits from being presented on contemporaneous hardware, old computers are usually either not available, very hard to maintain, or tend to fail when being used again after a long time of inactivity.

3. AN EMULATION KIOSK-SYSTEM

Recently, emulation frameworks have made great advances, in particular hiding technical complexity and by using web standards for delivery [6]. A technical emulation framework for public display has to be different from web-based emulation setups [2]. Running emulators on local machines (standard PCs) can be an interesting alternative for reading-room setups or museum displays, where cluster- or cloud-computing options are not suitable. For example, when running locally, emulators can provide much better response times then when run on remote infrastructure.

In an exhibition or reading room situation, the emulation setup typically needs to render only a single, specifically prepared artwork. To make such a system cost-efficient, a re-useable technical design is necessary, ideally only disk images and objects should be exchanged and a wide range of emulators should be supported. Furthermore, the technical system should be self-contained, such that it can be used without a network connection or similar additional requirements to the surrounding technical environment.

For exhibiting artworks with interactive components, the physical-technical context can be important, e.g. how the user interacts with the work. This is especially true for historical pieces. Hence, even though using a contemporary computer system (most importantly a contemporary CPU to run emulators) the environment should be flexible enough to support old (or old looking) peripherals or displays. Compared to a web-based presentation, local execution of emulators allows to connect peripherals, such as joystick or printers, different display options, e.g. CRT monitors, projectors etc., and supports an authentic user experience for applications such as games, software based art or net art by providing native fullscreen display and practically zero (input-)latency.

Finally, the system needs to be adapted for public display, in particular protecting the artwork and the installation from undesired manipulation. For interactive works, where a user is even asked to change the environment through usage, the system containing the artwork should be reset for every visitor so they encounter the work in its intended state. Especially for long-term exhibitions, the system needs to be simple to setup (e.g. simply power on a machine) and simple to fix, if the setup has crashed.

4. TECHNICAL DESIGN

In the course of the EMiL project[1] an emulation-based access framework for multimedia objects in libraries and museums has been developed. The EMiL system is an advancement of the bwFLA/EaaS system and aims at integrating with different catalogues and long-term preservation systems. The project consortium consists of the German National Library, the Bavarian State Library, Karlsruhe University of Art and Design and the University of Freiburg. The project is funded by the German Research Foundation (DFG). As a result of the EMiL project the EaaS system has been modified to run within a so-called live-system but using a common EMiL/EaaS codebase.

In particular, the EaaS system – originally developed as a cloud-enabled, web-based system – was adapted to make use of local hardware, in particular running emulators on locally available CPUs, use the machine's input devices as well as local available graphics hardware and display(s) attached. The live-system is tailored to run directly from a USB stick, such that it boots any standard computer system with no additional preparations or installation requirements. To achieve optimal hardware support, the EaaS live-system is derived from an Ubuntu live-system.[2] As the EMiL live-system was designed especially for library reading rooms or museum exhibitions, access to the Linux system is restricted by default and users can only select objects/environments to emulate and interact with the emulation UI.

Currently, the live-system contains two disk partitions. The first partition contains a read-only file system containing a ready-made installation of all necessary software components, i.e. emulators, the second partitions is writeable and contains by default two folders:

- `configs/` contains configuration files

- `image-archive/` an optional image-archive

While the first partition has a fixed size (currently about 1 GB), size and filesystem type of the second partition can be changed by the user, as long as the filesystem is supported by a current Linux kernel. For a demo setup[3], we choose the proprietary filesystem *exFAT*[4] in order to support virtual disk images larger than 4 GB and to be compatible with most major desktop operating systems.

[1] Multimedia Emulation, http://www.multimedia-emulation.de/

[2] https://help.ubuntu.com/community/LiveCD

[3] A sample USB image can be downloaded http://bw-fla.uni-freiburg.de/usb-demo.img We recommend to use a fast USB 3.0 stick, with at least 8 GB capacity.

[4] exFAT, https://en.wikipedia.org/wiki/ExFAT

4.1 System Configuration

The configuration directory (`configs`) contains

- `common/` common configuration, e.g. configure the inactivity timeout;

- `remote/` configuration of remote image- and object-archive;

- `local/` configuration of a local image- and object-archive;

- `X11/` custom `xorg.conf` options, only required for old CRT monitors without EDID[5] support or to use non-standard input peripherals.

The local configuration is always preferred by the system. Only if no image-archive (respectively object-archive) folder is present on the data partition, the remote configuration is used.

For debugging purposes, a *devmode* switch is available that disables safeguards against accessing the underlying Linux system, allowing full terminal access and access to the systems log-files while running on the target machine.

4.2 Object and Image Preparation

In order to setup a custom emulation environment, the user needs to provide a disk image, supported by one of the emulators, a digital artefact to be rendered using the disk image's operating system and installed software and metadata describing the complete setup.

The most simple way to produce and test a desired emulation setup is to use the EaaS web-based environment. Workflows can be used to adapt existing disk images, for instance, installing additional software, testing artefact's rendering performance or configuring the environment to autostart the rendering process. The result can then be downloaded and copied to the USB-drive's second partition (image-archive). Alternatively, emulation environment metadata can be edited manually.

Alternatively, the USB live-system integrates well with an existing EaaS environment, by configuring a remote image-archive and/or object archive. In this setting, emulators still run on the local CPU and are able to make use of locally attached peripherals, while content (images and/or objects) is served through the network. Currently, the USB live-system requires a cable network with enabled DCHP service to function. WiFi connections are not yet supported.

The installation can either be configured to boot directly into a specific environment by putting a file (`environment-id.txt`) into the top-level directory of the second partition. The file should contain only the ID of the environment to load. You can find the ID of an environment in its metadata.

Furthermore, the live-system supports a reading-room setup with web-based user interface, which allows users to choose an environment. This setting is default, if no specific environment is set via `environment-id.txt`. In this setting, the user is able to switch between a full screen view and a web-based view (*CRTL-ALT-F*). In the non-fullscreen mode, the user may have options to cite an environment, create a screenshot, change a medium, etc.

5. STAGING THE 20 YEARS ANNIVERSARY EXHIBITION OF MBCBFTW

In 2016, Olia Lialina's pioneer 1996 net art piece *My Boyfriend Came Back From The War* had its twentieth anniversary. Haus der elektronischen Künste (HeK) in Basel, Switzerland, ran a retrospective exhibition of this work, combined with versions created by other artists [4], running from January 20 to March 20 2016.

For the exhibition, four EMiL-based live-systems were used, running on standard Intel NUC Mini PCs. USB 3.0 thumbdrives were prepared containing one artwork each, as well as disk-images of the the required environments and operating systems – in this case Windows 98 and Windows XP, running Netscape 3 and Internet Explorer 6. The operating systems and browsers were set up using EaaS web workflows and then exported to the USB drives to auto-boot. The environments were configured to automatically start the browser containing the desired artwork, but were otherwise not locked down or limited in use. The audience was able to freely interact with the complete environment, for example using the Windows ”Start” menu to run Microsoft Paint, but any changes made to the environment were reset after five minutes of inactivity.

The network was set up to transparently connect to a locally running web archive server based on Rhizome's Webrecorder [5], so correct URLs would be displayed in the browsers even for long-defunct web sites. Since all web traffic was handled by the web archive, the gallery audience would not be able to leave the boundaries defined by curation. The web archive server was configured to only deliver dial-up speed connections.

Using standard adaptors, hardware contemporary with the works was connected to the modern Mini PCs: end-user grade 14” and 17” CRT screens, one 15” LCD screen, and ball mice delivered authentic historical input/output devices. Additionally, period computer cases were used as props, with cables placed as if the tower was connected to them (Fig. 1). The historic hardware was lent from the collection of the media restoration department at the Bern University of the Arts.

Since some of the CRT screens were unable to communicate their technical capabilities to the EMiL Linux kernel (either because they were built before I2C/Display Data channel was standardized, or they simply didn't support either interface), graphics modes and horizontal and vertical sync had to be forced via software settings. Since this poses a risk for damaging the monitors, the required modes had to be tried carefully. In general, when legacy CRT monitors are used, the risk of them failing is relatively high even when all settings are correct, just because of their age. It is advisable to have backup monitors in place for that case.

In other cases, if the data exchange between display and kernel works, and the requested graphics mode is much lower than the monitor's recommended default, a too-high resolution might be selected by the kernel, presenting the emulator's visuals centered on the screen instead of fullscreen. This is desirable to avoid image distortion when emulating a 4:3 display output to be shown on a 16:9 LCD screen for instance, but doesn't make sense on a low-end CRT. In this

[5] VESA Enhanced Extended Display Identification Data Standard, Video Electronics Standard Association (VESA), Feb. 9, 2000

case, again, the graphics mode has to be forced via software.

Legacy USB 1.0 peripherals, in this case the ball mice, which were connected to the Mini PCs via standard PS2-to-USB adapters, can cause a whole USB controller to switch back to the very slow USB 1.0 mode. As the EMiL emulation system boots from an external USB 3.0 thumbdrive, it is important to use computers with at least two separate USB controllers, so that the peripherals' bus is separated from disk access. In the case of the Intel NUC systems, the solution was to connect mice and keyboards on the front USB sockets and the thumbdrives on the back.

After these issues had been solved for the HeK exhibition, it was possible to send the emulation-based artworks on bootable USB thumbdrives via FedEx to other galleries and museums, who would again use standard PCs to exhibit them. These other institutions sourced legacy input/output devices (CRT screens and mice) from cheap Ebay offers or could used old equipment they still had in their possession.

From February 19 to March 30 2016, the exhibition was shown at MU in Eindhoven, The Netherlands. One additional work, a fast-paced action game, was added to be shown via EMiL running a MacOS 9 system with keyboard interaction. Thanks to the abstraction offered by bwFLA, this operating system was configured and exported to EMiL within the same web workflow as the Windows systems. The overall exhibition design was changed, but adhering to the same principles of legacy hardware and props. Glass table tops, exposing the Mini PCs running the emulators, were used to highlight the staging aspect.

Both the technical staff at HeK and MU have reported that the emulators have run with great stability throughout the whole exhibitions' times.

Parts of the exhibition were also shown at the exhibitions *Electronic Superhighway (2016 – 1966)* at the Whitechapel Gallery, London, UK [3] and *Mashup* at the Vancouver Art Gallery, Vancouver, Canada [1], using the exact same techniques.

6. CONCLUSION

The combination of emulated, fully accessible legacy environments, reduced network speed, web archives, legacy input/output output devices and props provided a rich, narrative techno-cultural context for the presented net artworks and defined a very practical definition of the artworks' boundaries.

EMiL has greatly normalized the work required for exhibiting complex net art in physical space.

Future work on the EMiL system will improve exporting and update mechanisms for emulators stored on local disks or thumbdrives, offer more local settings and simplify the setup process for graphic modes.

Digital art produced right now for current operating systems like Windows 10, Linux, or Mac OSX will be possible to be re-enacted in the future using the same techniques, since integration work for these and more legacy systems is ongoing.

7. REFERENCES

[1] D. Augaitis, B. Grenville, and S. Rebick, editors. *MashUp: The Birth of Modern Culture*. Vancouver Art Gallery, 2016.

[2] D. Espenschied, I. Valizada, O. Stobbe, T. Liebetraut, and K. Rechert. (re-)publication of preserved, interactive content âĂŞ theresa duncan cd-roms: Visionary videogames for girls. In *Proceedings of the 12th International Conference on Digital Preservation (iPres15)*, 2015.

[3] O. Kholeif, editor. *Electronic Superhighway*. Whitechapel Gallery, 2016.

[4] O. Lialina and S. Himmelsbach. *My Boyfriend Came Back From The War - Online Since 1996*. Christoph Merian Verlag, 2016.

[5] M. McKeehan. Symmetrical web archiving with webrecorder, a browser-based tool for digital social memory. http://ndsr.nycdigital.org/symmetrical-web-archiving-with-webrecorder-a-browser-based-tool-for-digital-social-memory-an-interview-with-ilya-kreymer/, 2016.

[6] D. S. Rosenthal. Emulation & virtualization as preservation strategies. https://mellon.org/resources/news/articles/emulation-virtualization-preservation-strategies/, 2015.

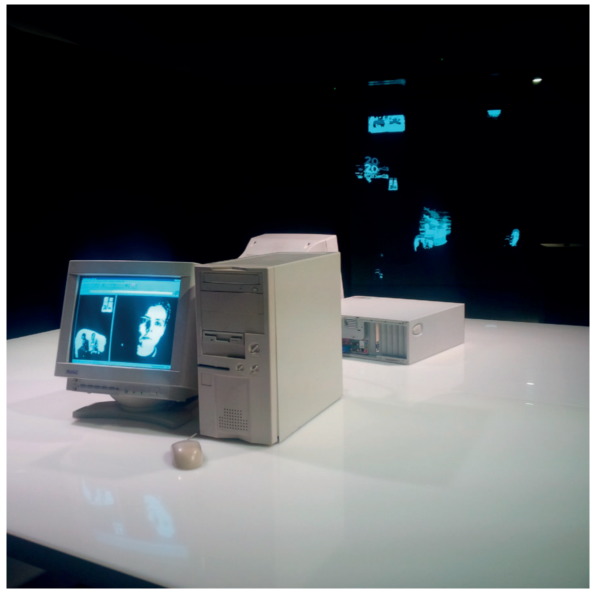


Figure 1: The MBCBFTW exhibition at HeK featuring mockup a "tower" computer case and a CRT screen connected to the Intel NUC (mounted concealed below the desk).

Project "The Digital City Revives"
A Case Study of Web Archaeology

Tjarda de Haan
Amsterdam Museum
Kalverstraat 92
1012 PH Amsterdam
+31 6 47132758
T.deHaan@amsterdammuseum.nl

ABSTRACT

Twenty-two years ago a city emerged from computers, modems and telephone cables. On 15 January 1994 De Digitale Stad (DDS; The Digital City) opened its virtual gates in Amsterdam. DDS, the first virtual city in the world, and made the internet (free) accessible for the first time to the general public in the Netherlands. But like many other cities in the world history, this city disappeared. In 2001 The Digital City, the website, was taken offline and perished as a virtual Atlantis. Although the digital (r)evolution has reshaped our lives dramatically in the last decades, our digital heritage, and especially the digital memory of the early web, is at risk of being lost. Or worse already gone. Time for the Amsterdam Museum and partners to act and start to safeguard our digital heritage. But, how to excavate The Digital City, a virtual Atlantis, and reconstruct it into a virtual Pompeii? In the case study of web archaeology we will try to answer the questions: how to excavate, reconstruct, present, preserve and sustainably store born-digital heritage and make it accessible to the future generations? [1]

Keywords

The Digital City, web archaeology, digital heritage, digital preservation, collaboration

1. INTRODUCTION

De Digitale Stad (DDS; The Digital City) is the oldest Dutch virtual community and played an important role in the internet history of Amsterdam and the Netherlands. For the first time internet was (free) accessible to general public in the Netherlands. DDS is an important historical source for the early years of the internet culture in the Netherlands. The virtual city and its inhabitants produced objects, ideas and traditions in new digital forms such as web pages, newsgroups, chat, audio and video. DDS was a testing ground, and operated at the cutting edge of creativity, information and communication technology and science. It was not only an experiment with computers, but an experiment with questions, problems and challenges posed by the emerging information and communication technology.

And things were moving fast on the electronic frontier. DDS followed the developments closely which resulted in several interfaces (cityscapes):

1. **DDS 1.0**: 15 January 1994; all information and communication was offered in the form of a text-based environment (command-line interface; MS-DOS, UNIX) in Bulletin Board System technology. The so called 'Free-Nets' in the United States and Canada where a major source of inspiration for the founders. Free-Nets were 'community networks', or 'virtual communities' developed and implemented by representatives from civil society ('grassroots movement'). The metaphor of the city was reflected in the organization of the interface. There was a post office (for email), public forums to meet other visitors, a town hall and a central station (the gateway to the internet).
2. **DDS 2.0**: 15 October 1994; entry to the World Wide Web with the first DDS website with a graphical interface and hyperlinks.
3. **DDS 3.0**: 10 June 1995; introduction of the interactive 'squares' interface, the basic framework of the city's structure. Each square had its own theme and character, and served as a meeting place for people interested in that particular theme. Visitors could find information and exchange ideas with each other. 'Inhabitants' could build their own 'house' (a web page), send and receive emails (worldwide!), participate in discussion groups, chat in cafes, take part in the 'Metro', vote etc.

The Dutch social network DDS proved to be very successful. During the first weeks in 1994 all modems were sold out in Amsterdam. In ten weeks' time 12.000 residents subscribed. There was 'congestion' at the digital gates. Over the years the DDS user base of 'inhabitants' was growing: in 1994 there were 12.000 users, in 1995: 33.000, 1997: 60.000, 1998: 80.000 and in 2000: 140.000. DDS attracted international interest for the design it had chosen: DDS used the metaphor of a city to structure the still relatively unknown internet and made the users into 'inhabitants' of the city.

But in 2001 The Digital City was taken offline and perished as a virtual Atlantis. Ten years later, in 2011, the Amsterdam Museum started the project re:DDS, the reconstruction of DDS. Not only to tell and show the story of this unique internet-historical monument of Amsterdam, but also –and more important- to raise awareness about the risk of the loss of our digital heritage. This was the beginning of our case study in web archaeology: how to excavate, reconstruct, preserve and sustainably store born-digital data to make it accessible to the future generations'?

2. CHALLENGES

The Digital City is a digital treasury from the early days of the web in the Netherlands. Our digital heritage is now at risk of being lost and with this we risk losing the early years of the internet in the Netherlands. We face many challenges on different levels.

2.1 Our Digital Heritage is Getting Lost!

"The world's digital heritage is at risk of being lost", the UNESCO wrote more than a decade ago in 2003, and "its preservation is an urgent issue of worldwide concern" [2]. The UNESCO acknowledged the historic value of our 'born digital' past and described it as "unique resources of human knowledge and expression". Only one year ago in 2015, Google's Vint Cerf warned, again, for a 'digital Dark Age': "Humanity's first steps into the digital world could be lost to future historians. We face a forgotten generation, or even a forgotten century. Our life, our memories (…) increasingly exist as bits of information - on our hard drives or in the cloud. But as technology moves on, they risk being lost in the wake of an accelerating digital revolution". [3]

2.2 Out of the Box

The acquiring and preservation of digitally created expressions of culture have different demands than the acquiring and preservation of physical objects. This is a new area for the heritage field in general and the Amsterdam Museum in particular. The museum has to cross boundaries, and get out of its comfort-zone to break new ground in dealing with digital heritage. To seek out new technologies, and new disciplines. To boldly dig what the museum has not dug before. How to dig up the lost hardware, software and data? And how to reconstruct a virtual city and create a representative version in which people can 'wander' through the different periods of DDS and experience the evolution of this unique city? The challenge is: can we –and how?- excavate and reconstruct The Digital City, from a virtual Atlantis to a virtual Pompeii?

2.3 Complexity of Born-digital Material

In crossing the boundaries and dealing with new (born-digital) material the museum encountered the following challenges:

- **Material:** born-digital material is complex and vulnerable and has various problems. Due to the rapid obsolescence of hardware and software and the vulnerability of digital files, data could be lost or become inaccessible. Another problem has to do with the authenticity. With born-digital objects it is no longer clear what belongs to the original object, and what has been added later. DDS is a complex information system with different applications. DDS was built on SUN systems, open source applications and self-written programs. How to preserve digital objects that are interactive, networked, process-oriented and context-dependent? And finally, important issues regarding to privacy, copyright and licensing form major questions.
- **Methods**: there is a difference between the (well known) web-harvesting and (relatively new) digital archaeology. Web-harvesting is the equivalent of taking a snapshot of a live object, while in our project we aim to recreate the object itself (or at least to create a representative version for people to access) from the 'dead' web. Since the data is no longer online, we first had to 'excavate' the digital artefacts. Fortunately, there are a some great internationally projects and initiatives that inspire us. With web-harvesting projects, such as The Wayback Machine and GeoCities, current data are harvested (scraped/mirrored) and displayed or visualized. In recent web archaeological projects, such as the restoration of the first website ever, info.cern.ch by CERN, and the project 'Digital Archaeology' of curator Jim Boulton, the original data and software are found and reconstructed, and shown on the original hardware or through emulation techniques.
- **Division of tasks:** who will take which responsibilities to retrieve, reconstruct, preserve and store born-digital heritage and make it accessible to the public? There is currently no central repository for tools (for example to read obsolete media), no comprehensive software library (for example to archive the old software, including Solaris, Windows, and MacOS), no central sustainable e-depot and infrastructure and there is a lack of web archaeological tools.
- **Historical (re)presentations and preservation strategies**: what are the (realistic) approaches into (re)presenting of historical data: 'historical true' or a 'quick and dirty'? How to preserve (and represent) historical digital-born data (migration, conversion, emulation, virtualization)?
- **Approach:** At present there is an alarming lack of awareness in the heritage field of the urgency (or funding) that our digital heritage is getting lost. We decided just to do it and act... (with lots of trial and error) and to start developing a roadmap to safeguard and preserve our born-digital past.

3. THE START

So in 2011 the Amsterdam Museum initiated the project re:DDS, the reconstruction of DDS, and started defining the objectives. To consider the project as a success the museum aimed to achieve the following goals:

- To give this unique (digital) heritage the place it deserves in the history of Amsterdam and to tell the story of DDS.
- To create a representative version of the internet-historical monument DDS in the Amsterdam Museum for people to visit and experience.
- To start a pilot in digital archaeology and share knowledge.
- To safeguard DDS in the collections of the heritage institutions for sustainable preservation.

To start the project the museum laid out the 're:DDS Roadmap':

1. Launch of the open history laboratory and a living virtual museum: http://hart.amsterdammuseum.nl/re-dds.
2. Bring Out Your Hardware & Finding Lost Data: crowdsourcing the archaeological remains (with The Grave Diggers Party as a kick-off event) and collect stories and memories.
3. The Rise of the Zombies: analyze and reconstruction.
4. Flight of the Zombies: presentation of the Lost & Found and a reconstruction.
5. Enlightenment: Let the Bytes Free!: Conclusions and evaluation.

Let's us take you back in time and share our first steps in web archaeology.

3.1 Crowdsourcing

First step was: let's find the data! There was no DDS archive, so the only chance of finding ánd piecing the DDS data together lay with people: old inhabitants, former DDS employees and volunteers. Web archaeologists Tjarda de Haan (guest e-curator of the Amsterdam Museum) and Paul Vogel (volunteer) had worked for DDS and had connections with former residents and (ex) DDS employees. To reach out to the crowd, calls and invitations were sent out through mailing lists, blogs and social media. Especially Twitter and Facebook proved to be indispensable tools to get for example old hardware which is no longer easily available.

3.2 Grave Diggers Party

On Friday, 13 May 2011 the Amsterdam Museum organized the 'Grave Diggers Party' with the Waag Society in Amsterdam. A party with a cause. A party to re-unite people and collect lost memories, both personal and digital.

We set up 'The Archaeological Site re:DDS', with a 'Working Space', with workstations to be able to collect the data on central storage, the 'Historical (e-)Depot'. Participants were able to dig in the Wayback Machine and store their excavations in the Historical (e-) Depot. Computers were used as excavators. Storage was used as buckets. UNIX commands and mice functioned as pades, pick-axe and trowels, scripts as metal detectors. Metadata were written down on 'find cards', so all lost and found artefacts (analog or digital) were documented: who brought in the material, were did it come from, how was it original used, what is the current state and who could we contact for more information. As well we set up a 'Museum Space', with 'Tourist Tours' in the 'Cabinet of Curiosities', where we show the lost and found artefacts (like old DDS servers, terminals, modem banks, tape robots, screenshots of images, manuals etc.).

So, we invited former residents, former employees and to DDS kindred souls to the Grave Diggers Party: "Help us dig up this unique city and be part of the first excavation work of the re:DDS!,". "Look at your attic and/or hard drives and bring all servers, modem banks, VT100 terminals, freezes, disks, scripts, zips, disks, floppies, tapes, backups, log files, videos, photos, screenshots and bring all your memories and stories you can find!".

Fifty enthusiastic (some international) cybernauts came along with full bags, hard drives and USB sticks to kick off the archaeological excavating. And after the party the Waag Society served for three weeks as an interactive archaeological site. In the 'Working Space' digital excavations were done, and the temporary exhibition was growing day by day.

During the 'Tourist Tours' people were interviewed and stimulated to share their stories and memories.

After three weeks of digging we found some great artefacts. De Waag Society found and donated two public terminals. The terminals were designed by Studio Stallinga in 1995. Residents of the city of Amsterdam who did not have a computer with a modem at home, could make free use of these public terminals in various public places in Amsterdam. The terminals were located among others in De Balie, the Amsterdam Museum, the public library and the city hall.

Former system administrators brought in discarded servers they rescued from the trash. We excavated servers with exotic names such as Alibaba, Shaman, Sarah and Alladin. Their pitiful status: cannibalized (robbed of components, the costs were very high so everything was always reused) or broken or the hard drives were wiped and reformatted. A former resident donated one of the first modem banks. Another former resident sent a specially made radio play for DDS in 1994, 'Station Het Oor' ('Station The Ear'). The play was made during the first six weeks after the opening of DDS. It was based on discussions and contributions of the first digital city dwellers. And we excavated thirty gigabytes of raw data, including backups of the squares, houses, projects. Furthermore we collected a huge amount of physical objects, like various manuals, magazines, photographs, videotapes and an original DDS mouse pad.

3.3 Freeze!

As cherry on the cake we excavated the most important and unique artefacts, namely the three DLT tapes, titled: 'Alibaba freeze', 'Shaman (FREEZ)' and 'dds freeze'. Together they form the 'freeze' of DDS of 1996. On 15 January 1996 DDS, the 'ten week experiment that got out of hand', existed for exactly two years. For the second anniversary of DDS, the pioneers of the social network sent a digital message in a bottle: "In the past two years, the Digital City has been focused on the future. This anniversary is a good time to take a moment to reflect on what happened the last two years in the city. High-profile discussions, large-scale urban expansion, people coming and going, friendships, loves and quarrels, the Digital City already has a long history. Three versions of the Digital City have been launched in two years. People come and go. Trends come and go. Houses are rigged and decorated. But where are all digital data kept? Who knows 5 years from now how DDS 3.0 looked like? The Digital City will be 'frozen' on Monday, January 15th at 18:00 o'clock. A snapshot with everything the city has to offer will be stored on tapes and will be hermetically sealed. The tapes with the data, along with a complete description of the programs and machines the city is run with and upon, will be deposited in an archive to study for archaeologists in a distant future". [4] The tapes however were never deposited, there was never any documentation made, and the tapes were more or less forgotten. Fortunately for us they were rediscovered a few weeks after the Grave Diggers Party.

The tapes came (of course) without the matching tape reader. After a frantic search an 'antique' tape reader was found in the National Library of the Netherlands in Den Hague, where it was used as ... a footstool. The Library, partner of the project, donated the tape reader to the project. In big excitement we started digging. But after two nights of reverse engineering the tape reader and the tape, we only found 70MB where we hoped to find 3GB. We were pretty disappointed. All sounds coming from the tape reader (tok, grrrrrr, beeeep) gave us the impression that containers of data were being flushed, and we thought the city had evaporated. We were puzzled. Had there ever been more data on the tape? Is this the 'freeze' we had hoped to find? What was broken, the tape reader or the tapes. Or both?

After our failed attempts to read the tapes and our bloodcurdling reverse engineering experiences (dismantling the reader and manually rewinding a tape), we called the Computer Museum of the University of Amsterdam for help. Would they be able to read the historical data of DDS of the tapes? Engineer Henk Peek started to excavate immediately. Only a few weeks later he mailed us a status report:

> I've excavated about 11 Gigabytes of data tapes.

> It looks very positive!

3.4 'The 23 Things of Web Archaeology'

We made such an enormous progress in, what we now call, our 'slow data project'. It was time for the next step: the reconstruction, bit by bit, byte for byte. But given the bias in our knowledge and the complexity, nature and scope of the matter, we needed the expertise of specialist institutions. Together with our allied partners we started to explore the (im)possibilities of the reconstruction of born-digital data. We started to document our findings in the form of '23 Things'. [5] [6] In every 'Thing' we aim to explain our bottlenecks, choices and solutions. In addition, each partner will describe the state of affairs in the field of its expertise, and will illustrate this, where possible, with recent and relevant examples. In this way we share our joint experience and knowledge and in doing so we hope to lower the threshold for future web archaeological projects.

Data are the new clay, scripts are the new shovels and the web is the youngest layer of clay that we mine. Web archaeology is a new direction in e-culture in which we excavate relatively new (born-digital) material, that has only recently been lost, with relatively new (digital) tools. Both matter and methods to excavate and reconstruct our digital past are very young and still developing.

In 'The 23 Things of Web Archaeology' we research the following issues:

- **Born-digital material**. What is born digital heritage? How does it work? How it is stored and used?
- **Excavate and reconstruction.** What are the current methods and techniques how to excavate and reconstruct born-digital material (the physical and digital remains and the context in which they are found)?
- **Make accessible and presentation**. Finally, we look at how we can interpret the remains and the context in which they are found and make it accessible.

4. TOOLS AND METHODS

To start with the reconstruction of the lost and found DDS data the Amsterdam Museum partnered up with the University of Amsterdam. We brought together Gerard Alberts, teacher and author of history of computing, the students and former DDS administrators and programmers. The university provided the domain specific knowledge, student power and … made the first effort of reconstructing the DDS applications and interfaces. In the next section we describe the work broadly, from dealing with the lost and found data of the Grave Diggers Party to the first reconstruction. A more detailed elaboration will be written in the (short) future.

4.1 Physical Media

As stated before getting the raw data of the lost and found media turned out to be quite a challenge. Old hard discs connectors were no longer in use (for example SCSI's). In the end we managed to find (crowdsourcing!) an old system, a SUN server that was still fully operational, which we used to extract the raw data from the different storage systems that we recovered.

The tapes proved to be a little harder. The earlier mentioned specialist Henk Peek had to take apart another tape drive to fix it. After manually cleaning the drive he managed to read out large sections of the tapes that could be reconstructed with SUN restore. In the end he managed to read 11 GB of data of the tapes.

4.2 eForensics

In extracting the data from the file images we used eForensic methodologies. This mostly comes down to using low level block copy tools such as 'dd' (a UNIX command) and extracting the files from the resulting file system images. Multiple readout passes of the low level were done. And the resulting images were check summed to ensure that low level media errors were not made. Outliers were discarded.

4.3 DeNISTing

An often used eDiscovery tool is deNISTing, from the National Institute of Standards and Technology (NIST). We planned to use deNISTing not only as a way to identify data files but also to use the hash set to automatically identify OS type, applications and the locations of directories in the total data set. However it turned out to be not as useful as expected. Since the NIST hash set is geared towards Windows.

4.4 Virtual Machines

One of the chosen methods of the project was to emulate all the systems in the original working state, through a combination of virtual machines and software defined networking. We planned to emulate the entire cluster. Sparc emulation however proved to be painfully slow and therefor unworkable.

4.5 Reconstructing

Compiling the C code, most of DDS code was written in C, proved to be an easier road than emulation. Even though include files and libraries had changed through time. The students managed to compile most of the programs in the end.

4.6 First Results

In the end the students, with support of former DDS employees, were able to A. reconstruct DDS3.0, the third and most known DDS interface, and B. build a replica using emulation, DDS4.0. An enormous achievement, and a huge step into the new (scientific) discipline web archaeology! Next step is to incorporate these results in the museum. DDS had already been included in the permanent collection of the museum, in a minimal version. The audience can see our first web archaeological excavations, the DDS 'avatars', take place behind the original DDS public terminal and watch television clips about DDS. In the nearby future we aim to enable people to interact with DDS, by taking a walk through the historical digital city, and 'experience' how the internet was at the beginning and how did it look like in the 20th century.

5. NEXT LEVEL

To enter the next level of our project we teamed up with the University of Amsterdam, Dutch Institute for Sound and Vision, Waag Society and started the project "The Digital City revives". With the joint forces of museums, innovators, creative industries, archives and scientists we will face our last major challenges: how to open up and sustainable store the DDS data into e-depots, how to contribute to a hands-on jurisprudence of privacy and copyright for future web archaeological projects, how to share our knowledge and lower the threshold for future web archaeological projects and how to present the DDS born-digital heritage in a museum context for future generations?

Our goals of our project "The Digital City revives" are:

- Reconstruct and preserve DDS.
- Provide insight into the (existing and new) processes, techniques and methods for born-digital material and the context in which they are found, to excavate and reconstruct.
- Ask attention to the danger of 'digital amnesia'.
- To provide museums and organizations with specialized knowledge about the reconstruction of born-digital heritage and lower the threshold for future web archaeological projects. Disseminating knowledge about new standards for archives on the storage of digital-born heritage in 'The 23 Things of Web Archaeology' and a 'DIY Handbook of Web Archaeology'.
- Make DDS data 'future-proof' in making the digital cultural heritage:
  - Visible (content): promoting (re)use of DDS.
  - Usable (connection): improving (re)use of DDS collection by making it available by linking and enriching data.
  - Preservable (services): maintain DDS sustainable and keep it accessible.

To be continued!

6. ACKNOWLEDGMENTS

We would like to thank our partners for their enormous contributions to our project. Our partners are (till now): Amsterdam City Archives, Creative Industries Fund NL, Digital Heritage Netherlands, Dutch Computer Heritage Foundation, Dutch Cultural Coalition for Digital Durability, International Institute of Social History, Karin Spaink, independent researcher, LIMA, National Coalition on Digital Preservation, National Library of the Netherlands, Netherlands Institute for Sound and Vision, old inhabitants, (ex) DDS employees, especially Paul Vogel and Michaël van Eeden, and DDS affiliated web archaeologists, University of Amsterdam, University of Amsterdam Computer Museum and Waag Society.

7. REFERENCES

[1] This article is partly based on an earlier article published in "Data Drift. Archiving Media and Data Art in the 21st Century" (October 2015), published by RIXC, LiepU MPLab, http://rixc.org/en/acoustsicspace/all/, by Tjarda de Haan and Paul Vogel.

[2] UNESCO published the Charter on the Preservation of Digital Heritage (October 2003). http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html

[3] http://www.theguardian.com/technology/2015/feb/13/google-boss-warns-forgotten-century-email-photos-vint-cerf

[4] http://web.archive.org/web/20100830120819/http://www.almedia.nl/DDS/Nieuws/freeze.html

[5] The 23 Things of Web archaeology is based on the Learning 2.0 – 23 Things, http://plcmcl2-things.blogspot.nl/, a program of Helene Blowers and inspired by the many sites of the 23 Things that followed.

[6] http://hart.amsterdammuseum.nl/23dingenwebarcheologie

Ellipse – Long-term and Permanent Protection and Accessibility of Geodata

Krystyna W. Ohnesorge
Swiss Federal Archives
Archivstrasse 24
3003 Bern, Switzerland
Tel. +41 58 464 58 27
krystyna.ohnesorge @
bar.admin.ch

Chiara Marciani
Swiss Federal Archives
Archivstrasse 24
3003 Bern, Switzerland
Tel. +41 58 462 48 70
chiara.marciani@
bar.admin.ch

Alain Mast
Swiss Federal Archives
Archivstrasse 24
3003 Bern, Switzerland
Tel. +41 58 463 5074
alain.mast@
bar.admin.ch

ABSTRACT

Archiving of geodata historically focused on methods of keeping digital geodata alive “almost forever”. Project Ellipse is a joint effort by the Swiss Federal Archives (SFA) and the Federal Office of Topography (swisstopo) running from 2011 to 2016. Its aim is to find a common solution for the archiving of geodata in order to implement the applicable legislation. Ellipse follows the entire archiving process chain: from the inventory and appraisal of geodata, to its submission to the digital archives and finally to the users, who expect geodata in a form that is authentic and accessible in a future technological environment.

Archiving of geodata is a complex task that demands intensive cooperation among all stakeholders. Despite our efforts, not all questions have been solved successfully. In this paper, we will report our findings and solutions as well as the obstacles we encountered during the course of Project Ellipse.

Keywords

Preservation of geodata, long-term availability, geoinformation system, Geo-SIP, Geo-Dossier.

1. INTRODUCTION

Geodata is (digital) information that identifies the geographical location and characteristics of natural or constructed features and boundaries on the earth’s surface, typically represented by points, lines, polygons, and other complex features (vector data) or pixels (raster data). These descriptive items are not understandable if they are not linked to geospatial reference data, for instance to topographical maps. The combination of spatial orientation with other thematic sets of geodata or with geospatial reference data creates geoinformation. In today’s modern society, geoinformation is the foundation for planning, measuring and decision-making at private and federal level. It is an integral part of state action and has to be preserved. Thus, Project Ellipse is developing a geodata archiving solution.

According to its remit, Ellipse shall achieve the following objectives:

- To develop an integrated solution for all geodata produced in the federal administration
- To achieve a worthwhile enhancement of the long-term availability of geodata and of archiving
- To allow geoinformation to be restored and interpreted from archived geodata at a later date

2. INFORMATION ABOUT THE PROJECT

The Archiving Act (ArchA)[1] and the Geoinformation Act (GeoIA)[2] require geodata produced in the federal administration to be preserved. For this reason, the Federal Office of Topography Swisstopo and the Swiss Federal Archives (SFA) were asked to develop a geodata archiving solution, in order to implement the applicable legislation. The scope of the project is limited to the geodata listed in the appendix to the GeoIO[3] (official geodata catalogue).

Between January 2011 and March 2013, Project Ellipse developed a concept for archiving official geodata. The concept describes the fundamentals for archiving geodata along the process steps production – geodata management – planning of conservation and archiving – acquisition (to the archive) – preservation – use. The main emphasis here was the collaboration between the producers of geodata and the Swiss Federal Archives. Based on the legislation on geoinformation, these institutions are required to mutually ensure the availability of geodata for the long term.

In the spring of 2013, the concept was approved by the SFA, swisstopo, as well as by the coordinating agency for federal geographical information (GCG), thereby initiating the implementation phase, which will be concluded by the end of 2016. Collaboration with cantons[4] and specialist organizations within both the archiving and the geoinformation communities will continue throughout this phase. The objectives were combined into four different work packages, each of which contains a range of tasks:

- **Work Package 1 – Conservation and archiving planning CAP**[5]: responsible for planning and conduction of appraisal of all geodata and for developing a tool to support this appraisal.
- **Work Package 2 – Formats and GeoSIP**: responsible for documentation and specification of archivable file formats and for developing a specification for a Geo Submission Information Package.
- **Work Package 3 – Access and Use**: responsible for creating and assuring access to geodatasets from archiving search platforms, and specifically for developing a link between the geometadata online catalogue (geocat.ch) and the archive metadata online catalogue (swiss-archives.ch).
- **Work Package 4 – Operational Organizations**: responsible for geo knowledge accumulation in the SFA and for developing and defining a second level support solution for end users.

Swisstopo assumed responsibility for work package WP1, the SFA leads the other three work packages. Below, the organisational structure of Project Ellipse is depicted:

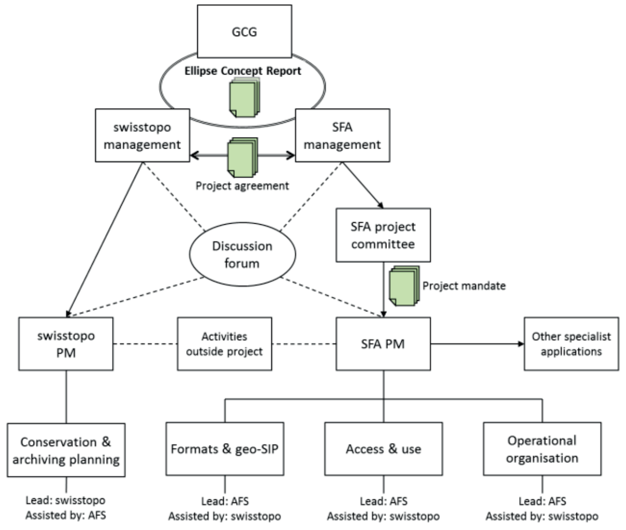


Figure 1 Project Ellipse organization

3. RESULTS

3.1 Conservation and Archiving Planning

A key issue in the management of geodata is to define what data must be available on which platform, to what purpose and for how long. In Switzerland, there is a legal framework for answering these questions, which distinguishes between *conservation* for a limited time at the authority responsible (long-term availability) and *archiving* for an unlimited time by the SFA.

The CAP includes all geodata sets and other spatial data sets of the Federation with the corresponding appraisal of the long-term availability and its value for archiving. The responsible offices of the Federal Administration conducted the first part of the appraisal. The SFA conducted the second part. The combined appraisal results identified the majority of geodata sets as archivable (313 out of 342 datasets). As most datasets are based on a decree, the need for their archiving was implicitly given. Furthermore, many datasets were built with a considerable effort and they promise broad potential for future use, adding another reason for their archiving. Of the remaining 29 datasets, 9 datasets were not considered for archiving, because they were products of data that was archived in a different, more suitable channel already. The remaining 20 datasets could not yet been appraised because either the datasets themselves or their juridical foundation was not yet completed.

3.2. Formats and GeoSIP

The work package *Formats and GeoSIP* concerns itself with file formats suited for the archival of geodata and with the definition of a submission information package for geodata.

With TIFF+Extended World File (TIFF+EWF.XML), a format for the archival of georeferenced image and graphic raster data has been defined and its specification has been published. The format consists of a baseline TIFF image and a simple XML sidecar file, which stores a minimum catalogue of attributes that permit the description with regard to space, time and content. Both files are linked by sharing a common filename that differs only in the extension.

The XML file contains ten attributes, of which six are the attributes that also make up a world file[6]. The remaining four attributes are:

- ReferenceSystem: Indicates the geographic reference system used in the form of a text reference in accordance with EPSG, for Switzerland “CH1903 / LV03” or “CH1903+ / LV95”.
- BeginTemporalExtent and EndTemporalExtent: Temporal extent of the content of the geodata or best possible approximation of the period in ISO 8601 format.
- ImageDescription: An optional free text to describe the image.

The decision to specify TIFF+EWF.XML instead of using GeoTIFF[7] was driven by the fact that there is currently very little use of GeoTIFFs in the Swiss Federal Administration. This means that the barrier of introducing a simple but new format like TIFF+EWF.XML is actually lower than introducing an established~~,~~ but more complex format such as GeoTIFF. Additionally, there is no standard set of metadata tags to be used in GeoTIFF which, in our opinion, further compromises long-term understandability of this format.

For archiving georeferenced vector data, no straightforward solution was found. In the Swiss Federal Administration, the products and thus the file formats of ESRI[8] (Shape and the Geodatabase family) are dominant. These proprietary formats however are not ideal candidates for long-term storage. Currently, there are only two candidate formats for archiving georeferenced vector data: On the international level there is GML, and on the national level, there is the Swiss format INTERLIS2[9]. It is important to note that the translation between any vector format is a challenging task, it is often ambiguous and therefore difficult to automate.

For lack of a better solution, INTERLIS2 will be named as format for archiving georeferenced vector data, since it is more widely used in the Swiss administration than GML. When submitting data in INTERLIS2, it will also be possible to additionally submit the most current ESRI format (of the same data), in the hope that this will facilitate the transition to other formats in the future.

[1] https://www.admin.ch/opc/de/classified-compilation/19994756/index.html

[2] https://www.admin.ch/opc/en/classified-compilation/20050726/index.html

[3] The appendix can be found at: https://www.admin.ch/opc/de/classified-compilation/20071088/index.html#app1

[4] Switzerland is divided into 26 administrative regions called *cantons*.

[5] http://www.geo.admin.ch/internet/geoportal/de/home/topics/archive_planning.html

[6] https://en.wikipedia.org/wiki/World_file

[7] http://trac.osgeo.org/geotiff/

[8] http://esri.com/

[9] http://www.interlis.ch/interlis1/description_d.php

The Swiss federal archives receive digital data for the archive encapsulated in SIPs (submission information packages). Initially in the project, it was planned to extend or adapt the SIP specification to assist archival and retrieval of georeferenced data. This plan will be postponed though (and thus pushed beyond the finishing date of Project Ellipse), as currently, several parallel undertakings to adapt or improve the SIP specification are on the way that must be streamlined to minimize impact on surrounding systems.

Instead, focus was shifted to the definition of a set of rules on how to organise geodata inside an SIP container in a structured way that takes into account the multi-dimensionality (time, space, layering) of geodata. The goal is to define a simple structure for storing geodata that is primarily understandable by humans, and secondarily aims toward automatic or semi-automatic machine-readability. The solution that will be proposed, while suitable for simple use cases, will be flexible enough to accommodate more complex use cases and geodata of different producers and systems. Additionally, it will be indifferent of formats and allow storing of primary data, necessary metadata and accompanying documentation. For that, we coined the term Geo-Dossier[10]. At the time of writing, a Geo-Dossier contains three first-level folders for storing documentation, models and primary data. It defines the mandatory splitting of data into subfolders if there are multiple views of the same data (e.g. multiple reference systems or multiple quality levels). It also allows for optional splitting of data into subfolders for arbitrary criteria (e.g. in thematic layers or spatial regions).

3.3 Access and Use

In this work package, requirements for the user interfaces were identified in order to enable the information retrieval system of the SFA to cope with geodata. Furthermore, geometadata required for archival purposes was selected and the linking between the access system run by swisstopo for long-term availability and the access system run by the SFA for archival was defined.

In the SFA, requirements for the user interfaces, for information retrieval and for search were defined in a series of workshops. A minimal set of requirements suitable for implementation within the currently operational access systems has been identified. Additionally, an extended set of requirements to be implemented in a possible future access system was defined. As the primary purpose of the access system of Swisstopo already is the handling of geodata, no additional functionality is needed there. In a next step, a set of geometadata to assist categorisation, search and information retrieval in the archival system of the SFA was selected:

- UUID: The identifier as it is used in geocat.ch, the metadata system of swisstopo (mandatory)
- Official geodata set ID: Another identifier that is mandatory for certain kind of Geodata
- Abstract: A short textual description of the data (optional)
- Preview: A thumbnail of a selected area of the geodata (mandatory)
- Georeference Data UUID: The identifier of a reference geodata set, if the actual geodata set is based on such (optional)
- Additional Georeference Data: If above UUID does not exist, a textual description of any reference geodata (optional)
- Geocategory: A categorisation based on the standard eCH-0166[11] (mandatory)
- Keywords: Keywords for describing the geodataset (optional)

By having the UUID as it appears in the long-term availability system of Swisstopo as a metadata, it is possible for both access systems to link to a dataset as it appears in the other system. Thus, users of one system can see that related datasets exist in the other system (usually the case if older data of a dataset has already been archived, while newer data is still only found in the long-term availability system).

3.4 Operational Organizations

The two activities of this work package consisted of appraising existing processes and identifying necessary changes in relation to acquisition and archival of geodata by the SFA, and of identifying and building up geo expertise inside the SFA.

The need for adjustment of the SFA-internal processes for data receiving and end user support proved to be negligible. Consequently, only a small internal workshop to convey basic geo knowledge for the SFA staff was conducted. Furthermore, a test run of the process of delivering geodata from the producer to the archival systems of the SFA was conducted between May and September 2016.

4. CONCLUSIONS

One of the key factors that have led to successful results was the broad guided dialogue we have had with all producers of geodata and the very efficient cooperation between all stakeholders. We maintained a high level of communication among SFA, swisstopo, the affected geodata producers and the Swiss geo community as a whole, which proved to be invaluable for the success of the project.

Undoubtedly the largest influencing factor in the project were the results of the CAP. Before the conclusion of the CAP, it was unknown how much, if any, geodata would be chosen to be archived and how long this data would remain in long-term availability before being sent to the SFA. Also unknown was the amount of data that the SFA had to expect and at what time intervals the data was to arrive, so it was difficult to judge the amount of automation that had to be built for ingesting.

The specification of a Geo-SIP was postponed, so that a more generic approach for a new and more flexible SIP specification can be developed, in which the accommodation of geodata will only be one part. This postponement however freed valuable resources for definition of the Geo-Dossier, a task that was not initially planned ~~for~~ but proved to be important.

The results of the work package Access and Use will influence the future of information retrieval in the SFA and will be a valuable input to the definition of the generic new SIP format.

The work package Operational Organisations has shown us that our processes are already flexible enough to accommodate themselves to various kinds of information, including geodata.

[10] Definition of the Geo-Dossier is still ongoing and results will be available at the end of 2016.

[11] Vgl. eCH-0166 Geokategorien, Version 1.1 (Minor Change) of the 23.09.2013, http://www.ech.ch

The main work that will go on past the conclusion of Project Ellipse at the end of 2016 is the

- definition of an archivable vector format and the
- definition of a new and more flexible SIP format which is better suited for various kinds of digital data, such as geodata or hypertext data.

Even with these activities still outstanding, we feel that Project Ellipse has successfully addressed the important aspects of archiving of geodata, and we are confident that with the current level of cooperation between all involved parties, the ongoing work can be adequately addressed.

5. REFERENCES

[1] Projektteam Ellipse, 2013, *Concept for the archiving of official geodata under federal legislation*, p.14. https://www.bar.admin.ch/dam/bar/en/dokumente/konzepte_und_weisungen/konzeptbericht_ellipse.pdf.download.pdf/concept_report_projectellipse.pdf

[2] Schweizerisches Bundesarchiv, 2016, *Bewertungsentscheid Geo(basis)daten des Bundes*. https://www.bar.admin.ch/dam/bar/de/dokumente/bewertungsentscheide/Geobasisdaten%20Bewertungsentscheid%202016.pdf.download.pdf/Bewertungsentscheid%20Geo(basis)daten%20des%20Bundes%20(Projekt%20Ellipse,%20AAP),%202016-02-19.pdf

[3] Schweizerisches Bundesarchiv, 2015, *Merkblatt Spezifikation archivtaugliches Geoformat für Bild- und Grafikrasterdaten TIFF + EWF.XML*. https://www.bar.admin.ch/dam/bar/de/dokumente/kundeninformation/Merkblatt%20TIFF%20EWF%20XML.pdf.download.pdf/Merkblatt%20TIFF%20EWF%20XML.pdf

Preserving Research Data: Linking Repositories and Archivematica

Jenny Mitcham, Julie Allinson
University of York
Heslington, York, UK, YO10 5DD
+ 44 (0) 1904 321170
jenny.mitcham@york.ac.uk
julie.allinson@york.ac.uk

Matthew Addis
Arkivum
R21 Langley Park Way
Chippenham, UK, SN15 1GE
+44 (0) 1249 405060
matthew.addis@arkivum.com

Christopher Awre, Richard Green, Simon Wilson
University of Hull
Hull, UK, HU6 7RX
+44 (0) 1482 465441
c.awre@hull.ac.uk
r.green@hull.ac.uk
s.wilson@hull.ac.uk

ABSTRACT

Applying digital preservation actions to research data is a practical step in carrying out research data management to its fullest extent and helps ensure this data remains usable in the future. This paper considers how repositories holding research data can link to an external third party tool, Archivematica, in order to carry out preservation actions as part of the data deposit workflow into the repository. We present experience from local use of Archivematica at the Universities of York and Hull in the Jisc Research Data Spring project "Filling the Digital Preservation Gap" as well as Archivematica as a shared service by Arkivum. A main focus across these use cases is a practical approach – parsimonious preservation – by using the Archivematica tools as they exist now whilst building a foundation for more comprehensive preservation strategies in the future. A key area of ongoing investigation covered by this presentation is dealing the with long tail of research data file formats, in particular how to best manage formats that are not immediately supported and need to be added to file registries such as PRONOM.

Keywords

Digital preservation; research data management; software as a service; repository integration; preservation workflows; file formats

1. AUDIENCE

This presentation is aimed at repository managers and related staff working to preserve digital content held within repositories. It is equally aimed at archivists and particularly digital archivists looking at ways to preserve both traditional archival material and other digital content collections, in particular research data.

2. BACKGROUND

Digital preservation should be seen as an integral part of Research Data Management (RDM). Research data is potentially very long lived, especially where it is irreplaceable and supports long running research studies, for example climate data, astronomy observations, and population surveys. This data will only remain usable if it undergoes active digital preservation to ensure that the applications of tomorrow can successfully find, retrieve, and understand the research data of today.

Digital Preservation is "the series of managed activities necessary to ensure continued access to digital materials for as long as necessary" where access is "continued, ongoing usability of a digital resource, retaining all qualities of authenticity, accuracy and functionality deemed to be essential for the purposes the digital material was created and/or acquired for" [1]. In the context of RDM, research data is kept to ensure that any research outputs based upon it are repeatable and verifiable[2] and also because research data has value through sharing so it can be reused and repurposed[3]. These underpin the ability to make research data openly available in a form that can be both used and trusted in the long-term.

Whilst digital preservation is clearly desirable, there can also be major challenges in its application, especially to diverse holdings such as University research outputs. This may come as a surprise given that there is no shortage of advice, guidelines and tools for digital preservation. There are dedicated organisations and resources available, including the Digital Preservation Coalition[4] and the Open Preservation Foundation[5]. There is a wide range of tools that can be used, for example as listed by COPTR[6]. There are increasingly well-defined processes for doing preservation, especially for data. Examples include workflows based on the functional model of the Open Archive Information System (OAIS)[7], which can be manifested in the policies/procedures of an organisation, for example the Archive Training Manual from the UK Data Archive[8] [9] Finally, there are also frameworks for assessing preservation maturity that provide a pathway for incremental progression along a preservation path, for example the NDSA Levels of Digital Preservation[10]. However, with this plethora of resources and tools comes complexity, cost, and a lot of choices that can lead to protracted timescales for getting preservation up and running within an institution.

[1] http://handbook.dpconline.org/glossary

[2] https://royalsociety.org/~/media/royal_society_content/policy/projects/sape/2012-06-20-saoe.pdf

[3] http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report,_Jan14_v1-04.pdf

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

[5] http://openpreservation.org/

[6] http://coptr.digipres.org/Main_Page

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

[8] http://www.data-archive.ac.uk/curate/archive-training-manual

[9] http://www.dcc.ac.uk/sites/default/files/documents/RDMF11/HERVE.pdf

3. SCOPE

This paper focuses on how Archivematica[11] can help institutions meet some of these challenges by delivering a practical solution to digital preservation. The paper focuses particularly on how this solution can be used for research data. This approach allows institutions to get started on the digital preservation ladder and then extend as their expertise and capabilities grow.

We show how Archivematica can provide a framework for digital preservation within an RDM infrastructure, including an example workflow for linking preservation with Institutional Repositories. The focus will be on the benefits of digital preservation and how it enables institutions to make their research data more accessible and useable over both the short and long terms.

The paper provides examples of how Archivematica is being applied in several contexts, and in particular will look at an ongoing project at the Universities of York and Hull which is actively investigating how Archivematica. This work is being undertaken as part of the Jisc "Filling the Digital Preservation Gap" project[12] on how Archivematica can be applied to research data [4][5], with a specific interest in workflow aspects[13] and how Archivematica can work with Institutional Repositories. Finally, the paper considers how institutions can lower the costs of adopting this solution thus enabling them to accelerate their preservation activities.

4. APPLYING ARCHIVEMATICA TO RESEARCH DATA PRESERVATION

Many of the benefits of using Archivematica stem from how it can be used to perform a technical audit that then underpins informed decisions on what to do about different types of research data. This is an essential part of 'parsimonious preservation'. This term was coined by Tim Gollins, Head of Digital Preservation at The National Archive in the UK [2],[3]. Being parsimonious means to 'get on and do' preservation in a simple and cost effective way that targets the immediate and real issues that digital content actually creates, rather than what the digital preservation community thinks might be problems in the future.

As University research data holdings diversify, digital content inexorably grows, budgets remain limited, and the benefits of easily accessible digital content become clear, there is never a more pressing time to apply the parsimonious approach. Archivematica provides a practical tool for parsimonious preservation, particularly in the areas of capturing and recording technical metadata within a preservation record (know what you have), and the safe storage (keep the bits safe) of data and metadata. Whilst focused on doing what can be done now, it also allows for additional tasks to be carried out in the future as required and as additional tools become available.

The approach of using Archivematica to inform decisions based on 'knowing what you have' can give an institution immediate visibility at time of deposit of whether the researcher's data is in a 'known' or 'unknown' format. For example, where a format is identified as 'unknown' (research data presents many uncommon formats) the institution can then work with the researcher on getting more information on the data format, securing software that can understand and render the data, or determining if the data format needs to be changed.

However, Archivematica is not purely about file format identification - there are a range of specific tools that could be used to identify files if this were the only requirement (for example, FIDO, FITS, Siegfried, JHOVE and DROID). File identification is just one of several micro-services initiated when a new dataset is transferred and ingested into the system[14]. Archivematica also performs other tasks such as checking the data for viruses (ClamAV), creating checksums (MD5, SHA1, SHA256 or SHA512), cleaning up file names, validating files (where appropriate tools are available) and carrying out format migrations (again where appropriate tools have been configured to do this). Currently, Archivematica identifies 720 file formats and has file format migration support for 123 of these using a wide range of tools (e.g. FFmpeg, ImageMagick, Ghostscript, Inkscape, and Convert). Whilst carrying out these tasks, Archivematica generates metadata describing the structure and technical characteristics of the dataset and this is packaged up (BagIt) and stored in the resulting Archival Information Package (AIP). These tools can all be used individually, but automation through Archivematica substantially reduces the time and effort involved in tool installation and then subsequent workflow automation.

5. PRESERVATION WORKFLOWS

The workflow shown below in Figure 1 is an example of how Archivematica could be integrated with an institutional repository and shows the use of Archivematica to assess and process content that has been deposited in the repository before it undergoes long-term archiving. Other options include using Archivematica to prepare content before deposit or using Archivematica to process content that has already been archived. These are discussed further in [1].

The workflow below shows how Archivematica can be used to ingest data created by a researcher as part of its deposit into the repository, this includes identifying which data is not in formats in Archivematica's Format Policy Register (FPR)[15], i.e. not in FIDO[16] or PRONOM[17], which can trigger an action to address this gap so that the data can be properly managed within the repository. In the first instance such management would allow the data to be presented correctly and links to relevant software made to enable engagement with it, adding value to the data in the repository over time rather than just holding it as a blob of uncharacterized data. In the longer term, knowledge of file formats within the repository also enables activities around Preservation Planning to take place, whether these consist of emulation or migration strategies.

[10] http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Levels_Archiving_2013.pdf

[11] https://www.archivematica.org

[12] http://www.york.ac.uk/borthwick/projects/archivematica/

[13] http://digital-archiving.blogspot.co.uk/2015/06/the-second-meeting-of-uk-archivematica.html

[14] https://wiki.archivematica.org/Micro-services

[15] https://wiki.archivematica.org/Administrator_manual_1.0#Format_Policy_Registry_.28FPR.29

[16] http://openpreservation.org/technology/products/fido/

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

Archivematica in this scenario might be operated by a digital archivist, a subject specific librarian or some other form of data expert, for example as part of technical support within a given research group or department (the Editor within Figure 1).

The workflow has the following steps (these correspond to the numbers in Figure 1).

1. Researcher uploads data files to the Repository in the normal way. Alternatively, this might be the institution's CRIS system.

2. The Researcher adds descriptive metadata.

3.The Editor reviews the Researcher's dataset, e.g. against minimum repository requirements.

4. As part of the review process, the data files are uploaded to Archivematica

5. Metadata is added if necessary. Archivematica and the tools it applies is used to in effect perform quality control on the dataset, e.g. to flag any files that don't have identified file types or any files that don't conform to their file format specification.

6. Archivematica generates an AIP, which is returned to the repository and stored in Repository Storage.

7. The Editor reviews whether processing in Archivematica was successful and that the dataset is correctly represented by the AIP. The Editor then approves the Researcher's submission and the Researcher is notified.

8. The AIP is sent to Archive Storage for long-term retention.

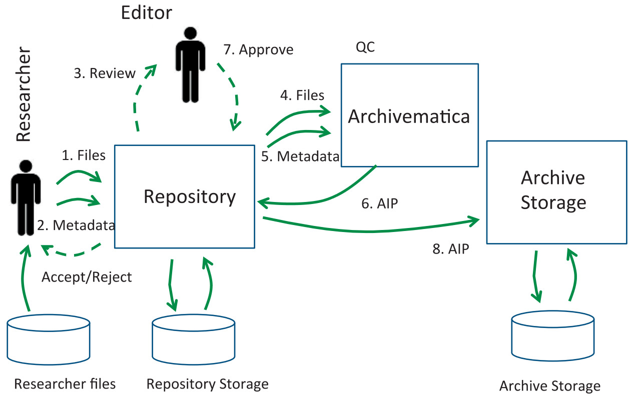


Figure 1 Example Archivematica workflow for research data preservation

Key to making Archivematica work with research data file formats is having a mechanism for reporting on unknown formats so that, in addition to local management, there is a way of adding value to the FPR and PRONOM by adding to the list of file format registered there and associated information about them. Work at York[18] (and reported in the Filling the Digital Preservation Gap Phase One report [4]) has highlighted the wide spectrum of research data file formats, and the long tail that preservation workflows will need to deal with over time. Though project work has started to address the problem through the addition of a small subset of research data formats[19] to PRONOM, this is clearly a problem that can only be addressed through wider collaboration and community engagement.

6. LOWERING THE COST OF USING ARCHIVEMATICA

There are several tangible benefits to using preservation tools such as Archivematica for research data, however, these benefits will only be realized if the associated costs are low. Whilst Archivematica itself is open source and freely available, this does not mean it is cost neutral. Costs include the resources and infrastructure needed for executing digital preservation, for example the start-up and ongoing costs of installing and running Archivematica pipelines. The human costs associated with the time taken to learn how to apply and use a preservation system should also be taken into account.

Whilst setting up their own implementations of Archivematica for the preservation of research data as part of the "Filling the Digital Preservation Gap" project, the Universities of York and Hull will be reviewing the costs of having done so. Certainly there have been cost savings in being able to work together on this project. Some decisions can be taken jointly and technical solutions around the integration of Archivematica with our repository systems can be shared. There is also a clear benefit to being able to talk to other Archivematica users. The UK Archivematica group has been a helpful and supportive community to share ideas and discuss solutions and there are undoubtedly benefits to working in an open way that enables us to learn from other people's mistakes and replicate their successes. Doing so can lead to cost savings in the longer term.

Another way that costs can be reduced for institutions is through use of a hosting model whereby a third-party provider delivers tools such as Archivematica as a service. The service provider handles the issues of setting-up, running, managing and supporting pipelines which allows the institution to focus on the archival and business decisions on what to preserve, how to preserve it, and the business case on why it should be preserved. It also addresses a current and significant issue that institutions have finding staff with the necessary skills in the installation, operation and maintenance of preservation software as well as their institution having the capacity to host and run this software on appropriate IT servers and storage systems.

Examples of communities that have started to establish common digital preservation platforms around Archivematica include the Ontario Council of University Libraries (OCUL) which has integrated Archivematica as part of the Dataverse research data publishing platform resulting in the ability to ingest data files and metadata from Dataverse into Archivematica for digital preservation purposes[20]. Also relevant in this context is the Council of Prairie and Pacific University Libraries (COPPUL)[21], Archivematica hosting and integration with DuraCloud[22], and Archivematica hosting and integration with Arkivum[23].

In the UK, recent developments under the Jisc Research Data Shared Service[24] provide another example of how institutions can work together on a shared approach to the preservation and management of research data. Archivematica has been selected as one of the digital preservation systems under this shared service and work is underway to ensure that it can be integrated with a number of repository solutions. As this new service is developed it provides a valuable opportunity for institutions to work together on their requirements and workflows and take advantage of a centrally hosted service.

[18] http://digital-archiving.blogspot.co.uk/2016/05/research-data-what-does-it-really-look.html

[19] http://digital-archiving.blogspot.co.uk/2016/07/new-research-data-file-formats-now.html

[20] http://www.ocul.on.ca/node/4316

[21] http://www.coppul.ca/archivematica

[22] http://duracloud.org/archivematica

[23] http://arkivum.com/blog/perpetua-digital-preservation/

[24]https://www.jisc.ac.uk/rd/projects/research-data-shared-service

Whilst using Archivematica as a hosted service from a third-party has many benefits, there are also several barriers to overcome. These include an assurance that an exit strategy is available in order to avoid lock-in to the hosting organization or to allow a continuity strategy that addresses the case where the service fails to be delivered. The use of open standards and data structures within Archivematica (for example PREMIS, METS, Bagit) is a key component of providing this assurance and allows migration to an alternative preservation service provider or in-house environment if needed.

7. CONCLUSION

Digital preservation of research data is an essential activity in ensuring that this data is accessible and usable in the future:

- Digital preservation has a valuable role to play in supporting the long-term availability and usability of research data, but it needs to be properly embedded into the research data management environment for these benefits to be realized.
- Digital preservation tools such as Archivematica can provide a quick way to get started with basic digital preservation whilst also providing a route for institutions to develop and apply more sophisticated techniques as their digital preservation maturity evolves.
- Doing preservation 'today' using Archivematica enables institutions to make practical parsimonious headway in the preservation of research data.
- Digital preservation tools are not currently able to recognize the range of file formats that researchers create. The digital preservation and research data community need to work together on improving the reach of file format registries and identification tools to help facilitate the preservation of research data.
- Archivematica is free but this does not mean implementation is cost neutral. There are ways of reducing these costs by sharing experiences and workflows, working together on integrations, and by taking advantage of the available hosting options.

8. REFERENCES

[1] Addis, M. 2015. RDM workflows and integrations for Higher Education institutions using hosted services, https://dx.doi.org/10.6084/m9.figshare.1476

[2] Gollins, T. 2009. Parsimonious preservation: preventing pointless processes!

http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation.pdf

[3] Gollins, T. (2012), Putting parsimonious preservation into practice, http://www.nationalarchives.gov.uk/documents/information-management/parsimonious-preservation-in-practice.pdf

[4] Mitcham, J., Awre, C., Allinson, J., Green, R., Wilson, S. 2015. Filling the digital preservation gap. A Jisc Research Data Spring project: Phase One report - July 2015, http://dx.doi.org/10.6084/m9.figshare.1481170

[5] Mitcham, J., Awre, C., Allinson, J., Green, R., Wilson, S. 2016 Filling the digital preservation gap. A Jisc Research Data Spring project: Phase Two report – February 2016, https://dx.doi.org/10.6084/m9.figshare.207

Implementing Automatic Digital Preservation for a Mass Digitization Workflow

Henrike Berthold
Saxon State and University Library
Dresden, Germany
+49 351 4677240
henrike.berthold@slub-dresden.de

Andreas Romeyke
Saxon State and University Library
Dresden, Germany
+49 351 4677216
andreas.romeyke@slub-dresden.de

Jörg Sachse
Saxon State and University Library
Dresden, Germany
+49 351 4677216
joerg.sachse@slub-dresden.de

Stefan Fritzsche
Technische Universität Dresden,
Germany
+49 351 46333212
stefan.fritzsche@tu-dresden.de

Sabine Krug
Saxon State and University Library
Dresden, Germany
+49 351 4677232
sabine.krug@slub-dresden.de

ABSTRACT

The Saxon State and University Library Dresden (SLUB) has built up its digital preservation system SLUBArchiv from 2012 to 2014. In January 2015, we launched the preservation workflow for digitized documents. This workflow extends the in-house mass digitization workflow, which is based on the software Kitodo.Production. In this paper, we describe the three major challenges we faced while extending our mass digitization workflow with an automatic preparation and ingest into our digital long-term preservation system and the solutions we found. These challenges have been

(1) validating and checking not only the target file format of the scanning process but also the constraints to it,

(2) handling updates of digital documents that have already been submitted to the digital long-term preservation system, and

(3) checking the integrity of the archived data as a whole in a comprehensive but affordable fashion.

Keywords

Digital preservation, Preservation strategies and workflows, Case studies and best practice, file formats, updates of archival information packages, bit-stream preservation, Saxon State and University Library Dresden, SLUB, SLUBArchiv

1. INTRODUCTION

SLUB has been digitizing its documents since 2007. The Dresden Digitization Center at SLUB is one of Germany's leading centers of mass digitization in the public sector. It produces 2 to 3 million scans a year. In addition, service providers digitize collections of other institutions as part of a digitization program of the federal state of Saxony in Germany. The software Kitodo.Production manages the digitization workflow i.e. the scanning process, the enrichment with structural and descriptive metadata and the export to the catalogue and to the digital collections presentation.

To preserve the resulting digital documents, SLUB has built up the digital preservation system SLUBArchiv. SLUBArchiv is based on the extendable preservation software Rosetta by ExLibris Corp. and complemented by a submission application for pre-ingest processing, an access application that prepares the preserved master data for reuse, and a storage layer that ensures the existence of three redundant copies of the data in the permanent storage and a backup of data in the processing and operational storage. Rosetta itself has been customized to SLUB's needs e.g. by plugins that have been developed in-house.

To complete SLUB's new digitization workflow (see Figure 1), an automatic pre-validation of the produced images has been added to check early if the requirements to the scanned files are met.

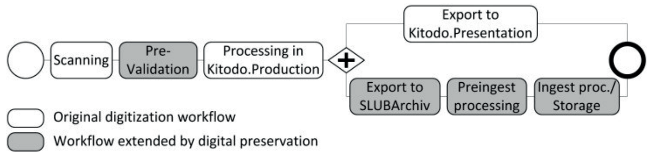

Scanning
Pre-Validation
Processing in Kitodo.Production
Export to Kitodo.Presentation
Export to SLUBArchiv
Preingest processing
Ingest proc./ Storage
Original digitization workflow
Workflow extended by digital preservation

Figure 1: Digitization workflow

In January 2015, the digitization workflow with digital preservation went live. In June 2015, SLUBArchiv has received the Data Seal of Approval [4].

During the development of the digital preservation for the digitization workflow, we faced a number of challenges. In this paper in section 3, we describe the three major challenges. The solutions are outlined in section 4. In section 2, we describe the mass digitization workflow in more detail, which is the setting for our challenges and solutions. In the last section, we summarize the future development goals.

2. MASS DIGITIZATION WORKFLOW

In SLUB's Digitization Center, a record is created in Kitodo.Production for each print document to be scanned. This record represents the digital document that corresponds to the print document. The document is then scanned on the appropriate scan device or, depending on the contents, possibly also on different scan devices (e.g. an enclosed map is scanned on a device that can handle large formats). All scans of a document are stored in the directory assigned to the digital document in Kitodo. When the scanning step is finished, checksums for all files are calculated and the processing starts. Descriptive metadata of the original print document are taken from the local or a remote library catalogue. Further descriptive and structural metadata are added. Finally, when the processing step is completed, the presentation data (in JPEG format) are exported to Kitodo.Presentation and the preservation data are exported to a transfer directory. The master data consist of one or more METS/MODS metadata files, zero or multiple ALTO XML files, and one or more master scans in the TIFF format. The steps of the workflow are visualized in Figure 1.

The transfer to the SLUBArchiv happens asynchronously. The submission application (i.e. our pre-ingest software) scans the transfer directory and processes the data of each newly arrived digital document. It checks completeness and integrity, transforms metadata from METS/MODS to METS/DC and converts the data to a SIP that is accepted by the software Rosetta. It then uses a Rosetta web service to initiate the ingest processing. During SIP processing, completeness and integrity is checked again. A plugin performs a virus check. The data format of each file is identified, validated and technical metadata are extracted. If SIP processing is successful, an AIP is built and stored in the permanent storage. The storage layer creates two more copies and manages them. The storage media used are hard disks in disk-based storage systems and LTO tapes in tape libraries. The archiving workflow is shown in Figure 2. It is fully automated.

Although the process seems to be simple, we faced a number of challenges. The most important challenges are described in detail in the next section.

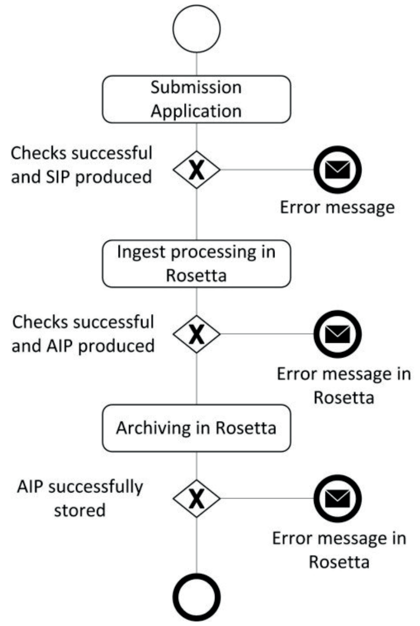


Figure 2. Archiving workflow

3. CHALLENGES

3.1 File Format Specification with Constraints

The specification of the data format TIFF 6.0 [1] specifies baseline TIFF and a number of extensions. During the ingest processing, the data format of each file is identified and the file is validated (i.e. it is checked whether the file is correct regarding to the file format specification). However, we have additional requirements regarding the digitized data. They have to be compliant with the guidelines specified by the Deutsche Forschungsgemeinschaft [2]. These guidelines specify values for technical parameters such as resolution or color depth. In addition, we have requirements that are important for digital long-term preservation [3]. One example is compression, which is allowed in TIFF 6.0 and encoded in tag 259 ("Compression"). So, to ensure robustness, we only accept uncompressed TIFF. Another example is the multipage feature, which allows for embedding multiple images in a single file (tag 297, "PageNumber"). To ensure that the metadata correspond to the image, we only allow one-page TIFF files.

3.2 Updates of Archival Information Packages

In contrast to the assumption that archived digital documents do not change, we have to deal with updates of digital documents. Reasons are manifold; some of them are:

- metadata are extended by a specific collection, e.g. due to an indexing project,
- a new representation is added, e.g. geo-referencing metadata,
- an image master has to be added or replaced, e.g. due to an error in the initial scanning process.

If a digital documents need to be changed, it is re-activated in Kitodo.Production. If the master data need to be changed, they are retrieved from Rosetta and reloaded into Kitodo. The digital document is then updated and exported again to Kitodo.Presentation and the SLUBArchiv.

3.3 Checking the Bit Stream of a Large Archive

SLUBArchiv currently preserves digital documents with a data volume of approx. 40TB. By the end of this year, the data volume will be approx. 100TB. We manage three copies of the data. Therefore, the total data volume is currently 120 TB, in 12/2016 it will be approx. 300TB. The storage layer of SLUBArchiv uses hierarchical storage management, in which large files are stored on tapes. An integrity check of all digital documents (and their three copies) is not feasible due to the time that is required to read all data from tape storage and check them. The estimated time effort is in the range of weeks. Since the complete restore cannot be done at once (because the disk storage cannot be held available for this amount of data), it would have to be organized in an incremental process. Such a process would stress the tapes. Therefore, we need a method to get reliable results without checking all data in the archive.

4. SOLUTIONS

4.1 File Format Specification with Constraints

Based on the code library libtiff [5], we have implemented the open-source tool checkit-tiff [6], which takes a human-readable configuration file and checks a tiff file against the specified configuration. It contains one rule per line (see example below). Each rule line has three entries:

- the ID number of a tiff tag,
- a string that specifies if the tag must be encoded in the file ("mandatory" or "optional"), and
- the feasible values as a single value "only(*value*)", a range "range(*from_value, to_value*)", a regular expression and some more options (see [6] for a complete list)

```
# Compression is not allowed
259; mandatory; only(1)
# XResolution
282; mandatory; range(300, 1200)
# YResolution
283; mandatory; range(300, 1200)
# Make i.e. name of the scanner
manufacturer
271; optional; regex("^[[:print:]]*$")
#PageNumber
297; optional; only(0,1)
```

Currently, we automatically check the files after the scanning step. If one or more files are not correct regarding the requirements, the digitization center or the external digitization service provider re-scans the files, replaces them in the directory of the digital document and the check is run again. This is to make sure that Kitodo processing only starts if all files are correct. Currently, the error rate is about 6%.

4.2 Updates of Archival Information Packages

We use the workflow software Kitodo.Production to produce and also to update digital documents. Hence, a specific document can be transferred multiple times to the SLUBArchiv – the first time is an ingest, all transfers after that are updates.

The transfer is done asynchronously. After the processing in Kitodo.Production, the files that belong to a document and that need to be preserved (i.e. master scan files, one or more METS/MODS files, ALTO XML files) are copied to a folder in the transfer directory.

The software Rosetta supports updates. It offers specific update functions through its ingest web service API. Rosetta can manage multiple versions of an AIP and creates a new version each time an AIP is updated. Older versions of digital objects remain stored and accessible for staff users.

The submission application (which takes the data that belong to a digital document and prepares a SIP) has to distinguish an initial ingest from an update. It uses the identifier of the digital document to check whether a digital document is currently processed by Rosetta or already archived. If the check is successful, it uses the update function, otherwise it uses the "normal" ingest function.

The Rosetta web service API provides functions to add a file, delete a file and replace a file. Using the checksums of the files, the submission application derives which file has been added, deleted or updated and applies the corresponding web service functions.

Currently, we are re-implementing the transfer from Kitodo.Production to the SLUBArchiv. We will make sure that all versions of a digital document that are copied to the transfer directory are archived. Since Kitodo.Production has no built-in versioning, we use the time of the export from Kitodo as our ordering criterion.

4.3 Checking the Bit Stream of a Large Archive

Each AIP is stored in three physical copies in two different locations, both of which are equipped with disk and tape storage systems.

Each AIP is stored in two storage pools - a primary and a secondary storage pool - of a clustered file system (IBM General Parallel File System, GPFS). The two storage pools are located at different locations. In these storage pools, large files are migrated to tape with IBM's Hierarchical Storage Management (HSM), which is an extension of the IBM Tivoli Storage Manager (TSM) software. The third copy of an AIP is a backup copy. TSM is used as backup software. Backup copies of new data in the GPFS-based permanent storage are made regularly (currently three times in 24 hours) and written to tape every day. All tape pools (i.e. HSM and backup tape pools) are protected by Logical Block Protection (LBP, a CRC checksum technology).

The integrity of archival copies is checked using two different methods.

4.3.1 Sample Method

Integrity of archival copies is checked yearly for a 1%-sample of all files. The sample of AIPs is produced automatically. Using Rosetta, a list of all files that belong to the selected AIPs is produced. The integrity of all three copies of these files is then checked automatically in the storage system. If an error is detected, corrupt files are replaced by correct copies and AIPs that are produced or updated on the same day are checked as well. Due to the applied storage policy, AIPs that are ingested on the same day are located in the same storage area. Depending on the results, the check is extended to a longer time period in which AIPs are stored in permanent storage. We have executed this check once. No integrity failures were found.

4.3.2 Pattern Method

The directory structure of the permanent storage can be controlled in Rosetta using a storage plugin. We have implemented a storage plugin that stores all files of an AIP in a single directory. These AIP directories are structured according to the year, month and day of the ingest. A file with a specified fixed bit pattern is stored daily in the directory of that specific day in the storage system. All these pattern-files are checked quarterly. Due to the specified bit pattern, single and multiple bit failures can be detected. If an error is identified, the data that are produced the same day are checked. Depending on the results, the check is extended to a longer time period in which AIPs are stored in permanent storage. We have executed this check once. No integrity failures were found.

5. CURRENT CHALLENGES

The SLUB Archive is developing towards new media types (digital video, audio, photographs and pdf documents), unified pre-ingest processing, and automation of processes (e.g. to perform tests of new software versions). Additionally, we currently conduct a pilot project of a digital preservation service for another Saxon institution.

6. REFERENCES

[1] Adobe Developers Association. 1992. *TIFF revision 6.0*. http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf

[2] Deutsche Forschungsgemeinschaft. 2013. *DFG Practical Guidelines on Digitisation*. http://www.dfg.de/formulare/12_151/12_151_en.pdf

[3] Rog, J. 2008. *Evaluating File Formats for Long-term Preservation*. https://www.kb.nl/sites/default/files/docs/KB_file_format_evaluation_method_27022008.pdf

[4] Saxon State and University Library Dresden. 2015. *Assessment of SLUBArchiv for the Data Seal of Approval*. https://assessment.datasealofapproval.org/assessment_178/seal/pdf/

[5] http://libtiff.maptools.org

[6] https://github.com/SLUB-digitalpreservation/checkit_tiff

Applied Interoperability in Digital Preservation: Solutions from the E-ARK Project

Kuldar Aas
National Archives of Estonia
J. Liivi 4
Tartu, 50409, Estonia
+372 7387 543
Kuldar.Aas@ra.ee

Andrew Wilson
University of Brighton
CRD, Grand Parade
Brighton, BN2 4AT, UK
+44 (0)1273 641 643
A.Wilson4@Brighton.ac.uk

Janet Delve
University of Brighton
CRD, Grand Parade
Brighton, BN2 4AT, UK
+44 (0)1273 641 620
J.Delve@Brighton.ac.uk

ABSTRACT

This paper describes the interoperability solutions which have been developed in the context of the E-ARK project. The project has, since February 2014, tackled the need for more interoperability and collaboration between preservation organizations. The solutions being developed include harmonized specifications for Submission, Archival and Dissemination Information Packages; and pre-ingest and access workflows. Furthermore, the specifications have been implemented using a range of software tools and piloted in real-life scenarios in various European archival institutions.

This paper provides a statement on the need for interoperability, and an overview of the necessary specifications and tools, and it calls for preservation organizations to continue collaboration beyond the lifetime of the E-ARK project.

Keywords

E-ARK; Interoperability; OAIS; SIP; AIP; DIP; pre-ingest; ingest; access; long-term preservation; digital archives.

1. INTRODUCTION

The adoption of increasingly sophisticated ICT in information creation and management has led to an exponential increase in the amount of data being created by / in a huge variety of tools / environments. Consequently, preservation organizations across the globe also need to ingest, preserve and offer reuse for growing amounts of data as well. When also taking into account the growing financial pressure which many organizations experience, we can conclude that there is a growing need for more efficiency and scalability in digital preservation. In more technical terms, institutions need to develop efficient guidelines and tools to support the export of data and metadata from source systems, produce or reuse metadata for preservation purposes, deliver information to the digital repository, ingest it, and finally provide relevant access services to appropriate end-users.

However, there is no single, widely understood and accepted approach on how valuable information should be transferred to digital repositories, preserved and accessed for the long-term [1]. In practice, existing approaches to archiving the same kinds of information are national or institutional, and differ in regard to their conceptual, technical and administrative underpinnings.

The European Commission has acknowledged the need for more standardized solutions in the area of long-term preservation and access, and has funded the E-ARK project[1] to address the problem. In co-operation with research institutions, national archival services and commercial systems providers, E-ARK is creating and piloting a pan-European methodology for digital archiving, synthesizing existing national and international best practices that will keep digital information authentic and usable over time. The methodology is being implemented in open pilots in various national contexts, using existing, near-to-market tools and services developed by project partners. This approach allows memory institutions and their clients to assess, in an operational context, the suitability of those state-of-the-art technologies.

The range of work being undertaken by E-ARK to achieve this objective is wide-ranging and ambitious, and more extensive than can be adequately described here. Accordingly, this paper will focus mainly on the Information Package specifications provided by the project, and introduce the range of tools which support these specifications.

2. NEED FOR INTEROPERABILITY

As mentioned above it is crucial to have more scalability and efficiency in archival processes. In particular, preservation organizations need to ensure that the data and metadata they receive and to which they offer access is formatted according to common and standardized principles. More specifically interoperability between source, preservation and reuse systems requires that:

- data and metadata are in standardized formats so their subsequent use is not inhibited by system differences;
- the data and metadata, and any other information required to use the data, are combined in a single conceptual package with all components being uniquely identified;
- the package contains enough information to allow validation both before and after transfer to a digital archive;
- the package is constructed in such a way that its information content can be understood in the long term without reference to external systems or standards.

In digital preservation terms this means that we need to come to a common agreement on the core technical and semantic principles of Information Packages (as defined in the OAIS Reference Model [2]). The main benefit of such standardization is that preservation organizations would be enabled to collaborate across institutional and legislative borders more effectively. Additionally, new opportunities would be opened for the reuse of tools which allow, in a standardized manner, the creation, identification, validation and processing of Information Packages. This, in turn, would reduce the effort

[1] http://www.eark-project.eu/

needed to maintain and develop bespoke local tools and ultimately save costs for any individual organization.

The E-ARK project has defined the standardization of Information Packages as its core activity. At the heart of this effort is the generalized E-ARK Common Specification for Information Packages (see 3.1 below). This specification is based on an extensive best-practice review of the available national and institutional specifications [1] and defines a common set of principles for how information being transferred and managed over time should be packaged to support interoperability and long-term access.

However, the Common Specification itself is not sufficient to achieve an adequate level of interoperability. In addition, the specific needs of pre-ingest, ingest, preservation and access processes need to be tackled. Accordingly, the project has also developed further, more detailed, specifications for Submission, Archival and Dissemination Information Packages. All of the specifications are based on the E-ARK Common Specification, but extend it with the specifics of the relevant processes (see 3.2 below).

The E-ARK Common Specification: SIP, AIP and DIP specifications can be called content agnostic as they allow the packaging of any data and metadata. However, to guarantee that the integrity and authenticity of information is not compromised, we need to also consider specific aspects related to the data in question as well as the environment from which it originates. For example, a typical real world records management system contains records arranged into aggregations, metadata relating to records and their relationships to other entities, a business classification scheme, a set of retention and disposal schedules, user access controls and definitions, a search engine and so on. All these data, metadata and environmental components, which make up a specific and complete information package, must be transferred together with the data in a way that the integrity, authenticity and understandability of the whole package are maintained. To allow for interoperability on such a fine-grained level, E-ARK has implemented the concept of Content Information Types (see 3.3 below). The Content Information Types provide a regime for specifying in detail the precise metadata, data, documentation, and system-level issues relevant for a particular type of Content Information, ultimately extending the scope of the Common Specification itself.

3. E-ARK SPECIFICATIONS

In this section we explain some of the details of the E-ARK Information Package specifications which are mentioned above.

3.1 Common Specification for Information Packages

The backbone of archival interoperability in E-ARK is provided by the so-called Common Specification for Information Packages [3]. The OAIS compliant specification is built on the requirements presented above and provides a unified set of rules for packaging any data and metadata into a single conceptual package which can be seamlessly transferred between systems, preserved and reused in the long term. The core of the common specification is a definition of an Information Package structure (Figure 1).

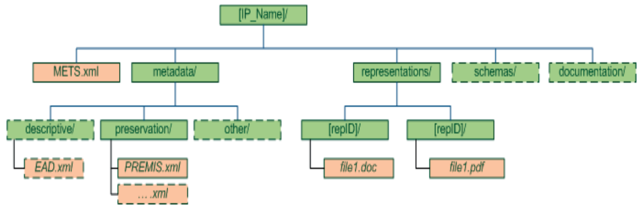


Figure 1: Basic E-ARK Information Package Structure

The structure allows for the separated inclusion of any metadata, data, relevant schemas and documentation into the package. Furthermore the metadata in the package can be divided into descriptive (metadata needed to find and understand the data), preservation (metadata needed to ensure the integrity and authenticity of data, metadata and the whole package) and other (any other metadata which is deemed relevant by the source system or the preservation organization).

A specific feature of the data component is that it can contain one or more representations of a single intellectual entity. The Common Specification allows also a single representation to include only the data of the specific representation or even duplicate the whole structure (Figure 2). Exploiting the last option allows implementers to differentiate between the package as a whole and a specific representation. For example, organizations can include generic descriptive metadata into the root metadata folder and at the same time keep detailed preservation metadata only within respective representations. Also, this offers the possibility of describing new emulation environments as a separate representation, thereby not endangering the integrity and authenticity of the original data and metadata. However, such splitting of metadata between the package and representations is purely optional within the Common Specification.

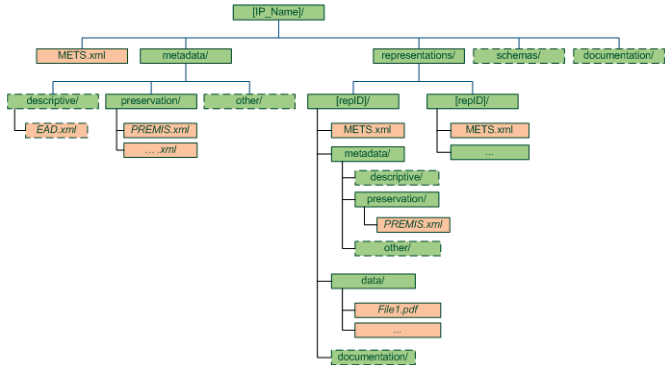


Figure 2: Full E-ARK Information Package Structure

Lastly, to ensure that the whole package can be understood and reused in the long term, users have the possibility of making the package self-sustaining by including any relevant schemas and documentation which might not be available externally in the future.

As well as the mandated folder structure, the information package folder must include a mandatory core metadata file named "METS.xml", which includes the information needed to identify and describe the structure of the package itself and the rest of its constituent components. As the name indicates the file must follow the widely recognized METS standard[2]. The METS.xml file needs also to be present in all representations in the case where the full folder structure is being used (Figure 2). The METS metadata serves the main purposes of:

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

- identifying the package and its components in a persistent and unique way;
- providing a standardized overview of all components of the package;
- connecting relevant pieces of data and metadata to each other.

In short, the METS metadata is the main tool and driver for interoperability that allows everything inside the information package to be validated according to commonly accepted rules.

In comparison to the METS standard itself, the Common Specification imposes a few additional requirements to be followed. One key requirement is the availability of a specific structural map (METS <structMap> element) which must describe the data, metadata and other components of the package. Again, this requirement is the key towards allowing different developers to create interoperable tools for validating and checking the integrity of Information Packages. Furthermore, the Common Specification provides some additional rules, for example a specific regime for linking to external metadata from the METS file, the availability of IDs etc. For full details of the Common Specification METS profile please consult the full Common Specification document [3].

To support the scalability of Information Packages, the Common Specification allows also for the splitting of intellectual entities across multiple physical packages or for the creation of Archival Information Collections (AICs). This can be achieved by formatting individual representations or parts of representations as a stand-alone Common Specification package and creating a "grandfather" IP which provides references to all of the components. The only requirement for both the components and the grandfather IP is the availability of the METS.xml file, created according to the regime defined in the Common Specification.

3.2 SIP, AIP and DIP Specifications

As mentioned above, the Common Specification provides a set of core requirements which are both process and content agnostic.

To cover for the needs of specific archival processes (pre-ingest, ingest, preservation and access) the E-ARK project has developed separate Submission, Archival and Dissemination Information Package specifications. While all of these specifications follow the rules of the Common Specification, they also widen its scope via the addition of specific details.

The E-ARK Submission Information Package specification [4] concentrates on the details of the pre-ingest and ingest processes. As such it provides additional possibilities for describing a submission agreement in the package, adding further details about the transfer process (i.e. sender and receiver), etc.

The E-ARK Archival Information Package specification [5] concentrates mainly on the need for authenticity. As such it describes multiple possibilities for adding new representations in the original Information Package by either including these in the original package or formatting them as new Common Specification packages. Furthermore, the E-ARK Archival Information Package specification makes special arrangements for keeping the original submitted Information Package intact throughout any preservation actions.

The E-ARK Dissemination Information Package [6] concentrates on the details of access needs. For example, it makes special provisions for the inclusion of specific Representation Information as well as order related information. It can include also an "event log" which can be used for proving the authenticity of the package, even when the original submission itself is not included in the package and is not provided to the user.

3.3 Content Information Type Specifications

As discussed above, an Information Package can contain any type of data and metadata. However, the types of data files, their structural relationships, and metadata elements vary for different Content Information types. For example, metadata produced by a specific business system will variously be intended to support different aspects of descriptive, structural, administrative, technical, preservation, provenance and rights functions.

The METS standard used in the E-ARK Common Specification does not offer one, single structure in which content type specific metadata could be stored as a whole. In order to efficiently use metadata to support archival functions, the Common Specification defines separate METS sections as containers for the various metadata functions, such as the METS header for package management, the <dmdSec> for EAD[3] and other descriptive metadata standards, and the <amdSec> for preservation (PREMIS[4]), technical and other functions. In order to use the submitted metadata, the content type specific metadata elements need to be mapped to those METS sections and implemented using agreed standards. To ensure interoperability on such a detailed content-specific level, complementary metadata profiles are needed for key Content Information types to define how the submitted content-specific metadata should be mapped to the E-ARK Common Specification structure.

To meet this need, the E-ARK Common Specification allows for the creation of additional Content Information Type Specifications. In effect, these detailed specifications can detail the specific requirements for package metadata, data structure, and relations between data and metadata. Essentially anybody is welcome to set up new Content Information Type Specifications as long as these do not conflict with the requirements presented in the Common Specification.

The E-ARK project itself has developed two such specifications:

- SMURF [7] (Semantically Marked Up Record Format) specification, which details the archiving of data and metadata from Electronic Records Management Systems (the specification is semantically based on the MoReq2010[5] standard) or for simple file-system based (SFSB) records (specification based on the EAD standard). The SMURF profile specifies in particular how to archive the necessary elements of an ERMS system, including the classification scheme, aggregations and classes, disposal schedules, and user access controls.
- Relational Database Profile which is based on the SIARD format [8]. SIARD is an open format developed by the Swiss Federal Archives. The format is designed for archiving relational databases in a vendor-neutral form. The format proposes a common standard for describing core elements of the live DBMS: data; structure; stored procedures; triggers; views; and queries.

4. TOOLS

As mentioned above, the E-ARK specifications are primarily intended to lead interoperable tool development. To validate the

[3] https://www.loc.gov/ead/

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

[5] http://www.moreq.info/index.php/specification

applicability of the specifications in real life scenarios the project has committed to providing a set of software tools that automate the creation and processing of Information Packages created according to the specifications. Also, as is a typical convention for EC-funded projects, E-ARK has committed to providing all the tools as open-source, freely available for the whole community to use and participate in developing. The project has not committed itself to developing a single tool for any specification but instead aims to provide a set of tools for the same task within the archival workflow (Figure 3).

For example, the basic SIP creation task can be handled by four quite different tools - ESS Tools for Producers (ETP)[6], RODA-in[7], Universal Archiving Module (UAM)[8] and E-ARK Web[9]. All of these tools implement specific features which make them suitable for different users. The ETP allows for the setup of complex ingest profiles and is therefore suitable for larger organizations; RODA-in excels in the package creation of "loose data" (for example when archiving a whole hard drive at once); UAM is specifically able to deal with data originating from Electronic Records Management Systems and E-ARK Web is a lightweight web-based environment for creating Information Packages manually.

However, all of these tools create Information Packages according to the E-ARK Common Specification. Therefore, they can be used for creating packages not only for transfer to a specific national or institutional repository but to ANY repository supporting the Common Specification as an input format.

This example also illustrates very well the aim of the E-ARK project. Once a common agreement is achieved on core technical principles, organizations will be able to select their tools out of a set of different options, instead of being obliged to use a fixed choice that is then linked to a fixed standard.

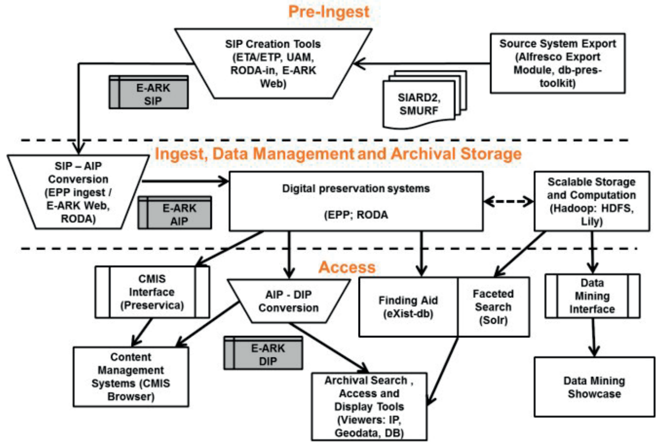

Pre-Ingest
SIP Creation Tools (ETA/ETP, UAM, RODA-in, E-ARK Web)
Source System Export (Alfresco Export Module, db-pres-toolkit)
E-ARK SIP
SIARD2, SMURF
Ingest, Data Management and Archival Storage
SIP – AIP Conversion (EPP ingest / E-ARK Web, RODA)
E-ARK AIP
Digital preservation systems
(EPP; RODA
Scalable Storage and Computation (Hadoop: HDFS, Lily)
Access
CMIS Interface (Preservica)
AIP - DIP Conversion
Finding Aid (eXist-db)
Faceted Search (Solr)
Data Mining Interface
Content Management Systems (CMIS Browser)
E-ARK DIP
Archival Search , Access and Display Tools (Viewers: IP, Geodata, DB)
Data Mining Showcase

Figure 3: Overview of the E-ARK toolset

To demonstrate this in practice, the E-ARK project is running dedicated pilot installations at seven sites in six European countries where selected E-ARK tools are deployed alongside already available infrastructures. The pilots run between May and October 2016, with results published by early 2017.

5. CONCLUSIONS

Ongoing access to information is a *sine qua non* of the modern world. But long-term access to and re-use of information depends, crucially, on ensuring the reliable and error free movement of information between their original environments and the digital archives. Additionally, the movement of information between different environments may occur many times during its lifespan and requires robust interoperability between those environments.

Thus, an approach for ensuring that digital information can be easily and consistently transferred between systems with all their characteristics and components intact is an urgent requirement for memory institutions. With its Common Specification, E-ARK has developed a coordinated approach to, and agreement on, standardized methods for packaging and sending information between systems, which is OAIS compliant. With its range of accompanying tools, the E-ARK approach has the potential to simplify and make consistent the currently diverse approaches to solving the issue of information transfer.

However, such standardization needs also to be carried on beyond the lifetime of the E-ARK and we are making every effort to ensure that the work of the project is also acknowledged, continued and broadened by the whole digital preservation community, not only the project partners. This is effectuated by the broad dissemination of the project from the outset via the partners DLM Forum[10] and the DPC[11], and also the practical involvement of three highly-involved advisory boards. The project outputs will be sustained long-term by the DLM Forum.

6. ACKNOWLEDGMENTS

This paper is presenting work which has been undertaken within E-ARK, an EC-funded pilot action project in the Competiveness and Innovation Programme 2007-2013, Grant Agreement no. 620998 under the Policy Support Programme.

We would like to acknowledge the work on the Common Specification by Karin Bredenberg of the Swedish National Archives and seconded for a period to E-ARK partner ES Solutions; to Dr. Angela Dappert (now at the British Library) on the SMURF specification; to Dr. Krystyna Ohnesorge's team at the Swiss Federal Archives on the SIARD 2.0 specification.

7. REFERENCES

[1] E-ARK Project. D3.1 E-ARK Report on Available Best Practices (2014). Available from http://eark-project.com/resources/project-deliverables/6-d31-e-ark-report-on-available-best-practices

[2] CCSDS. Open Archival Information System Reference Model (OAIS), 2012. Available from http://public.ccsds.org/publications/archive/650x0m2.pdf

[3] E-ARK Project. Introduction to the Common Specification for Information Packages in the E-ARK project. Available from http://www.eark-project.com/resources/specificationdocs/50-draftcommonspec-1

[4] E-ARK Project. D3.3 E-ARK SIP Pilot Specification. Available from http://eark-project.com/resources/project-deliverables/51-d33pilotspec

[6] http://etp.essarch.org/

[7] http://rodain.roda-community.org/

[8] http://www.arhiiv.ee/en/universal-archiving-module/

[9] https://github.com/eark-project/earkweb

[10] http://www.dlmforum.eu/

[11] http://www.dpconline.org/

[5] E-ARK Project. D4.3 E-ARK AIP Pilot Specification. Available from http://eark-project.com/resources/project-deliverables/53-d43earkaipspec-1

[6] E-ARK Project. D5.3 E-ARK DIP Pilot Specification. Available from http://www.eark-project.com/resources/project-deliverables/61-d53-pilot-dip-specification

[7] E-ARK Project. D3.3 SMURF – the Semantically Marked Up Record Format – Profile. Available from http://eark-project.com/resources/project-deliverables/52-d33smurf

[8] Swiss Federal Archives, E-ARK Project. eCH-0165 SIARD Format Specification 2.0 (Draft). Available from http://eark-project.com/resources/specificationdocs/32-specification-for-siard-format-v20

Copyright in Cultural Heritage Institutions
Snapshot of the Situation in Switzerland from a national library perspective

Andrea Ruth Schreiber
Swiss National Library
Hallwylstrasse 15
3003 Bern
+41 58 465 30 84
andrearuth.schreiber@nb.admin.ch

ABSTRACT

This paper outlines some of the most relevant copyright obstacles libraries in the cultural heritage institutions sector currently face, when trying to fulfill their mission in the digital context. For each of the four main activities – collecting, cataloguing, making available and preservation – the essential copyright issues will be outlined against the background of the legal situation in Switzerland. Where possible, short references to a broader copyright context and the laws of other countries will be given.

Particular emphasis will be placed on cataloguing and making available documents as the current ongoing Swiss copyright revision contains some innovative approaches: a catalogue privilege as well as new regulations for the handling of orphan works and mass digitization. Concerning collecting and preserving, at least some relevant questions in relation to copyright will be posed in order to maybe launch further discussions.

Keywords

Digital Heritage Institution ; Library ; Copyright ; Switzerland

1. INTRODUCTION

As the formulation of law usually lags behind technical demands and perspectives it becomes increasingly problematic for libraries to fulfil their mission in the digital context. They must deal with a multiplicity of legal problems especially relating to copyright and data protection, which generates difficulties on two levels, both in everyday practice as well as in the strategic development of an institution. This paper focuses on the matter of copyright.

Copyright issues are becoming more and more important in most libraries, in particular when it comes to the digital context. The variety of problems differ according to the type of the library in question, for example:

- A general public library will most likely deal with questions related to lending rights or the business models for acquiring and lending e-books that are mainly novels and nonfiction.
- Scientific libraries will focus on the development of new publishing models of scientific content, so the keywords for them are journal rates, licensing models and open access.
- National libraries treat their publications as cultural assets. This generates special copyright questions when cataloguing their collections and making them available to the public as well as preserving them, ideally forever.

Discussing the whole range of copyright issues occurring in different types of library would go far beyond the scope of this short paper. A decision had to be made, so the following explanations focus mainly on *copyright issues in libraries which serve primarily as cultural heritage institutions* [called CHI in the following]. Of course an elaborate presentation and a full enumeration of all the copyright difficulties in CHIs is still not possible within a few pages. But with regard to the ongoing copyright revision in Switzerland[1] (as likewise in numerous countries around the world), it will outline some of the most serious copyright issues relating to collecting, cataloguing making accessible and preserving cultural heritage.

As Switzerland is not a member state of the European Union, it has in some manner a wider range of possibilities to solve specific copyright problems. For example the famous EU-guideline for orphan works has not been implemented in Swiss copyright law. The draft regulations of Swiss copyright outlined below include some alternative and creative approaches regarding the needs of CHIs.

Nevertheless it would of course not make sense to presuppose that Switzerland is completely independent in formulating new copyright rules. Switzerland has also signed the major international copyright treaties[2] and, considering the internet as a global phenomenon, we need common solutions or at least approaches which once will function in a cross-border way.

The author of the present paper is not a lawyer but has been working in the copyright field in the Swiss National Library for several years. The following explanations have therefore not being developed in a 'legalistic' way. They rather refer to practical questions which evolve out of a practical librarian point of view but which are determined by the legal context.

[1] see https://www.ige.ch/en/copyright/modernisation-of-copyright.html [16/06/2016]

[2] especially
- Berne Convention for the Protection of Literary and Artistic Works, 1979, Berne.
- International Convention for the Protection of Performers, Producers of Phonograms and Broadcasting Organizations, 1961, Rome.
- Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS), 1994, Marrakesh.

2. COLLECTING DIGITAL RESOURCES

When acquiring digital works for their collections, CHIs face already fundamental copyright issues, especially if these works are digitally born and only accessible or available on the internet.

On the federal level in Switzerland there is no legal deposit law, neither for analogue nor for digital publications. Similarly to many other European countries, Swiss copyright law does not include an elaborated fair-use, as known in the United States for instance. Together, these two lacks make it particularly difficult for CHIs to integrate digital works in their collections: for example, e-books which are produced by a publisher in the traditional model usually cannot be bought, but only licensed. Of course licensing models are not an appropriate way of 'collecting' with the aim of long-term preservation of digital resources. Moreover, digital products such as e-books are often endowed with technical rights management devices in order to exclude unauthorized uses. Swiss copyright, like copyright law in every country which implemented the WIPO Copyright Treaty[3] and the WIPO Performance and Phonograms Treaty[4] (both signed in 1996), forbids circumventing such technical protection measures.[5] Besides, there are also publishers which refuse to supply e-books to libraries at all, fearing that this could reduce retail sales.[6]

But collecting digital heritage is not only difficult in a commercial context as is usually the case with e-books or e-journals. Non-commercial websites or blogs, for example, normally combine a lot of copyright protected works of different kinds and from diverse copyright holders. From a copyright view, all of these should give their permission in order that a CHI is legally allowed to collect and preserve their copyright protected content.

Licensing models as provided by Creative Commons and other non-profit organizations provide an interesting approach to improve the situation. If more and more creators put free licenses on their work, efforts for rights clearing processes can be reduced. However, the effectiveness of such licensing systems will depend on how many creators are actually going to use them. It can be expected, that especially rights' holders of works made for commercial use will rarely use free licenses.

Over all, to really enable cultural heritage institutions to include a multiplicity of copyright protected, digitally born content in their collections, a legal basis is needed, that:

► allows libraries to get digital publications such as e-books in a long-term and definitive way, without digital protection measures. This could be achieved by the introduction of an appropriate legal deposit or through a specific copyright exception for libraries working in the cultural heritage sector.

► enables CHIs to legally collect non-commercial digital works available on the internet such as websites in simplifying right clearing processes. A corresponding legal deposit (as some countries already have) is probably the most obvious solution.

3. CATALOGUING AND ACCESS

CHIs always have created inventories for their collections in order to make the works included searchable. Nowadays such inventories exist usually as electronic catalogues accessible on the internet. Their records consist of descriptive metadata such as title, name of the creator, extent, date of origin, etc. Of course, it would be far more effective and therefore user-friendly to make available online as much useful information as possible about the catalogued works. This could prevent that users be obliged to come on-site (or in case of libraries maybe order) to consult their works of interest – as was the case over the past centuries – to check, if the chosen works meet their expectations.

Both cataloguing and making available copyright protected works on the internet may create copyright problems, which must be taken particularly seriously. On the one hand, it is essential for CHIs to respect the authors' rights of the works in their collection. On the other hand, providing informative and user-friendly catalogues and making works collected available online are probably the most obvious missions of CHIs, which often are funded by public money. Therefore, new copyright solutions are needed which allow both a fair balance to the rights' holders as well as practical possibilities for CHIs to make their cultural heritage collections available in the digital age.

3.1 Catalogues

The present copyright draft in Switzerland proposes a completely new and unique regulation, formulated as a legal limitation for CHIs. Accordingly CHIs would be allowed to improve their online catalogues by enriching them with extracts of the catalogued works as long as this does not affect the rights and legitimate interest of the rights' holders. Regarding publications for instance, CHIs would be authorized to include covers as well as tables and abstracts of contents. Concerning pictorial works such as paintings or photographs, it should be possible to include small pictures in low quality (to avoid their re-use). In case of audio- and audiovisual materials, CHIs would be allowed to integrate limited short excerpts in their online catalogues. According to the drafted regulation, the use of works within this kind of catalogue enrichment neither demands rights clearing processes nor remuneration.

The implementation of this new 'catalogue privilege' would mean a real progress for CHIs, as they would be able to provide more informative, attractive and high quality inventories with a deeper insight into the described works. Since users would get a lot more information when searching online, they could decide more easily if it would be useful for them to consult a certain work or not.

There would be still some questions remaining relating to the practical implementation of the catalogue privilege. For example, how to ensure that a 'small picture in low quality' will still be good enough to generate the required added value for the users? At least the main content or main point of an image should be made recognizable. Otherwise it would not make sense at all to include it in the catalogue. But how to provide legal security for CHIs? Definition of pixels and size would be maybe the easiest but surely not the most reasonable way. Would it then be wiser to prescribe certain image formats for example – and if yes what kind?

As many CHIs hold millions of photographs in their collections, a more detailed definition of specifications regarding to the practical implementation of the drafted catalogue privilege would be crucial. And surely such specifications would highly influence strategic decisions on bigger retro-digitization projects in CHIs.

[3] see http://www.wipo.int/treaties/en/ip/wct/ [11/04/2016]

[4] see http://www.wipo.int/treaties/en/ip/wppt/ [11/04/2016]

[5] Girsperger, 2007, 118.

[6] see IFLA principles for library e-lending http://www.ifla.org/node/7418 [19/04/2016]

3.2 Obtaining Rights

Libraries' collections are numerous and heterogeneous. They usually contain a mass of works of different kinds and by many different creators. Accordingly the number of diverse copyright holders is immense. The Swiss National Library for example holds around 4.5 million publications, approximately 1.5 million photographs and graphic works, around 45'000 posters, several hundred literary and artistic archives and also a large number of audio works. The biggest part of the collection consists of works which have been created during the 20th century up until today. Most of them are still in copyright and thus cannot be made legally available on the internet without the permission of the rights' holders.

A high percentage of 20th century works are orphaned, which means that the rights' holders are not known anymore or cannot be traced and contacted in order to give the necessary permissions. Other than the catalogue issue described above, difficulties relating to orphan works and individual rights clearance in CHIs are better known, at least within the most relevant communities. Pamela Samuelson, copyright expert and one of the key speakers at iPRES 2015, also discussed this problem and spoke about related legal developments in the United States.

In Switzerland the copyright draft takes a dual approach.

3.2.1 Single Use of Orphan Works

A new draft regulation would allow the use of an orphan work from a CHI's collection after remuneration has been paid and therefore permission obtained from the representing collecting society. Furthermore, some research about the rights' holder must have be carried out, to ensure that the work in question is in fact orphaned. Contrary to the EU-directive[7] the Swiss approach does not define what sources must be searched. This can be seen as both an advantage and a disadvantage from the CHI's view: the EU-directive has been criticized a lot by CHIs for its very intricate and therefore unrealistic 'diligent search'.[8] But the non-definition of the necessary research leaves institutional users in legal uncertainty and private users (who according the Swiss draft would be also allowed to use orphan works) helpless, as they would hardly be experienced in undertaking effective research about rights' holders.

A clear advantage of the Swiss draft over the European regulation is the partial inclusion of integrated works. If an orphan work includes further orphan works, the rights for these must not be cleared separately, as long as integrated works do not determine the character of the work in question. This means for example, if a book of poetry which is orphaned also includes some orphaned illustrations, the latter must not undergo the same copyright procedure as the book itself.

Contrary to the EU-Directive the Swiss draft does not include any kind of a register of works which have once been declared as orphaned. The absence of such an inventory is the main disadvantage over the EU-approach, as one cannot trace for which works research about the rights' holders has already been made. This could result in multiple searches for rights' holders for one and the same work.

To summarize, the Swiss approach could be useful for individual uses of orphan works, if users and collecting societies work together. But it is surely no help when using a large number of orphan works, for example on the internet, as the efforts and costs for rights clearing processes would be far too high.

3.2.2 Mass Digitization

One of the main problems arising when digitizing large heterogeneous collections from the 20th century (which usually include a lot of orphan works) is that the rights' holders of works created in the first half of the 20th century are often not member of a collecting society, as those were only founded during the 20th century. As a result, all these rights should be cleared individually, which is of course impossible. To solve this problem, the Swiss copyright draft includes a very open version of the Scandinavian extended collective licensing-model [ECL]. According to the ECL-model, collecting societies are enabled to represent not only their members but also non-members, as long as their works correspond to the characteristic types of works represented by the appropriate collecting society.

The ECL outlined in the Swiss copyright draft is very general. Unlike some other European countries which already use the ECL, there is no limitation concerning the range or duration of use under the ECL. Both can be freely negotiated between the user and the appropriate collecting society. Furthermore everybody (not only privileged institutions) would be allowed to negotiate contracts with collecting societies based on the ECL-model. And contrary to the United States the drafted ECL in Switzerland is not seen as a time limited trial.

The actual introduction of such an ECL would of course be crucial for the strategic planning of large digitization projects of collections which include orphaned and other copyright protected works. Without a comparable tool, the results of such projects could not be made available at all. Hence the argumentation for conducting such projects would miss the most attractive key point.

The uptake of the ECL in Swiss law is therefore welcome, not only from the collecting societies' point of view but also from the perspective of larger heritage institutions which could afford such major digitization projects.

Again, success and practicability – or even abuse – of the drafted regulation will depend on the quality of interaction and negotiation between the different stakeholders. From the CHIs' perspective, the negotiated contracts should also undergo checks by an independent instance.

3.3 Text and Data Mining [TDM]

As in other ongoing copyright revisions, text and data mining [TDM] is being also discussed in Switzerland. The current drafted regulation allows text and data mining only for scientific purposes and against remuneration to the collecting societies.

From the user's point of view, remuneration is disputable as the largest part of the data to be mined is usually raw data, which is not copyright protected anyway. Moreover more and more publishers sell regular licenses for text- and datamining of their products. An additional remuneration would therefore go far beyond the objective. These arguments were also crucial in the United Kingdom, where in 2014 a new exception for text and data mining has been introduced – without remuneration. Furthermore, the

[7] DIRECTIVE 2012/28/EU OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 25 October 2012 on certain permitted uses of orphan works.

[8] see EBLIDA, 2015, 4.

limitation on 'scientific use' could be seen as problematic, especially as long as there is no particular definition of the term 'scientific'. In relation to the missing definition many further questions and uncertainties could arise.

4. LONG-TERM PRESERVATION

Swiss copyright law allows retro-digitization as well as the use of protected works for purposes of archiving and long-term preservation. As long as the works in question will not be made available, retro-digitization and other copyright relevant uses in relation to digital long-term preservation (regarding migration or in the context of emulation) are therefore permitted.

From a strategic point of view, it becomes more and more attractive for CHIs to move long-term preservation into the cloud in order to benefit from lower storage costs as well as to profit from the sustainability of cloud systems. Given the fact that most CHIs do not have sufficient resources to build a cloud on their own and to host the content by themselves, the outsourcing of archiving and long-term preservation of digital material becomes an interesting opportunity This raises additional legal questions not only in relation to data protection (which will not be treated here) but also in the context of copyright.

4.1 Transmission of Legal Privileges?

The above-mentioned, already existing regulation in Swiss copyright concerning the use of copyright protected works within the scope of archiving and preservation, is limited to special types of institutions such as libraries, educational institutions, museums and archives[9]. Thus the question arises whether these privileged institutions can legally outsource their long-term preservation to a third party such as a commercial company for example, which as such does not profit from the outlined archive and preservation privilege.

At least as long as the rented service could be subsumed under 'Infrastructure-as-a-Service' [IaaS], this is legally possible, supposing that the supplier provides only storage services and does not process the data, as well as access to the data is protected and only possible for the data provider.[10]

4.2 Territoriality of Copyright and Clouds?

As in most other countries, Swiss international private law recognizes the 'Schutzlandprinzip' (lex loci protectionis). Accordingly, Swiss law applies to violations that occur in Switzerland and foreign law applies to violations occurring abroad. This 'Schutzlandprinzip' corresponds to the general territoriality of copyright. In consequence, copyright violations will be judged according the law of the country in which the violation has taken place.[11] Out of this evolve further questions, especially regarding to outsourcing long-term preservation to cloud systems. While for example Swiss copyright includes the mentioned exception for long-term preservation under certain circumstances, other countries do not have this kind of regulation in their copyright law.

As the cloud user usually doesn't know in which countries the data will be stored and hosted, he can hardly make sure that the necessary migrations and other copyright relevant uses of the protected material according to long-time preservation are legal in the different countries in which the corresponding servers are located.

5. SUMMING UP

CHIs face a wide range of copyright questions, uncertainties and problems when trying to legally fulfill their main tasks: collecting, cataloguing, making available and preserving works from the cultural heritage sector.

Some important ambiguities relating to collecting and making available by now seem to have been taken up by wider communities. Accordingly various legislative processes in a number of countries do integrate first approaches in order to enhance the actual situation for CHIs. Unfortunately, this does not mean that the different attempts we have seen so far would provide real and practical solutions. But at least a start has been made – in Switzerland as well as in many other countries.

At the same time new questions relating to new techniques arise, for example in relation to outsourcing long-time preservation or the use of clouds. One of the biggest challenges is surely dealing with the territoriality of copyright – not only in the case of preservation but also of cross border uses when making available copyright protected digital heritage collections on the World Wide Web.

Making the relevant communities realize the range of copyright problems in CHIs, as well as searching for solutions together with other stakeholders, especially the rights' holders of the works in their collections, is a big task for CHIs nowadays. They must make sure that they won't be forgotten in the diverse ongoing political and economic discussions about dealing with advantages and disadvantages of technical progress or new internet business models. It could finally even be crucial for CHIs to make decision makers aware of the present copyright issues, in order to promote legal approaches which will allow CHIs to continue their cultural mission in the name and on behalf of society and the public itself.

6. ACKNOWLEDGMENTS

My thanks goes to the Swiss National Library, which encouraged and enabled me to make this ~~a~~ contribution to iPRES 2016.

7. REFERENCES

[1] Beranek Zanon, N. and de la Cruz Böhringer, C. 2013. Urheberrechtliche Beurteilung von IaaS- (und XaaS)- Cloud-Diensten für die betriebliche Nutzung gemäss Art. 19 URG. In *sic!* (11/2013), Schulthess Juristische Medien AG, Zürich. 663-681.

[2] EBLIDA (2015). *EBLIDA comments and voting recommendations on the amendments presented at the JURI committee in the draft report on the implementation of directive 2001/29/EC of the European Parliament and of the council of 22 may 2001 on the harmonization of certain aspects of copyright and related rights in the information society*. http://www.eblida.org/News/2015/EBLIDA_JURI-InfoSoc-Opinion-Voting-Recommendation2015-June.pdf [16/05/2015].

[3] Fehlbaum, P. and Lattmann, S. S. (2009). Schranken und anwendbares Recht bei Urheberrechtsverletzungen auf dem Internet. In *sic!* (05/2009), Schulthess Juristische Medien AG, Zürich, 370-381.

[4] Girsberger, M. 2007. *Schutz von technischen Massnahmen im Urheberrecht. Die WIPO Internetabkommen in den Vereinigten Staaten, der Europäischen Union und der Schweiz*. Stämpfli Verlag AG, Bern.

[9] see article 24 paragraph 1^{bis} of the Swiss copyright act.

[10] see Beranek Zanon, de la Cruz Böhringer, 2013.

[11] see Fehlbaum, Lattmann, 2009, 381.

The Ties that Bind - On the Impact of Losing a Consortium Member in a Cooperatively Operated Digital Preservation System

Michelle Lindlar
TIB Leibniz Information Centre for
Science and Technology
Welfengarten 1B
30168 Hannover, Germany
+49 511 762 19826
michelle.lindlar@tib.eu

ABSTRACT

Cooperatively operated digital preservation systems offer institutions of varying size the chance to actively participate in digital preservation. In current times of budget cuts they are also a valuable asset to larger memory institutions. While the benefits of cooperatively operated systems have been discussed before, the risks associated with a consortial solution have not been analyzed in detail.

TIB hosts the Goportis Digital Archive which is used by two large national subject libraries as well as by TIB itself. As the host of this comparatively small preservation network, TIB has started to analyze the particular risk which losing a consortium member poses to the overall system operation. This paper presents the current status of this work-in-progress and highlights two areas: risk factors associated with cost and risk factors associated with the content. While the paper is strictly written from the viewpoint of the consortial leader/ host of this specific network, the underlying processes shall be beneficial to other cooperatively operated digital preservation systems.

Keywords

Digital Preservation Services; Digital Preservation Networks; Consortial Systems; Risk Assessment; Exit Scenario.

1. INTRODUCTION

Digital preservation is per definition a risky business – or as Corrado and Moulaison put it: “Ultimately, digital preservation is an exercise in risk management” [1]. Much research has gone into the assessment of risks associated with digital preservation [2]: risks associated with file formats [3][4], risks associated with specific business cases and the application of risk assessment methodologies such as SPOT (Simple Property-Oriented Threat) or SWOT (Strengths, Weaknesses, Opportunities, Threats) to repositories [5],[6]. The focus of these assessments is either content-driven, i.e. focusing on problems specific to certain collections, or institutional repository driven, i.e. considering an institutional repository as a closed ecosystem.

Simultaneously, with institutions facing budget cuts, a growing number of institutions are turning to digital preservation networks, joint system implementations and preservation services such as DPN (Digital Preservation Network) or the MetaArchive Cooperative.

Despite the wide adoption of preservation networks, many supporting digital preservation actions maintain an institutional repository fixed view. Certification processes, for instance, such as the Data Seal of Approval, the nestor seal or the TRAC Audit process usually audit the participating institutions separately, even if they are participating in a single central digital preservation repository. This leads to a distinct blind spot regarding consortial management. A central question not answered by this approach is the following: what happens, if an institution leaves the consortia? While it can be assumed that the impact highly depends on the overall size of the consortia, the risks associated with an institution leaving touch on different areas and should be evaluated carefully.

Preservation networks as well as collaboratively operated systems range from *small networks* of 2-5 institutions, such as that of the National Library of New Zealand and Archives New Zealand in the National Digital Heritage Archive [7], to *mid-sized networks* of 6-20 institutions which are often found at the regional or state level, such as DA-NRW, the digital archive of North-Rhine-Westphalia in Germany[1], to *large national or international networks* with over 20 institutions, such as DPN – the Digital Preservation Network[2] – with over 60 members. More importantly, networks and collaborations differ in modi operandi regarding overall available preservation levels as well as responsibilities. In order to adequately assess the impact a leaving institution has on a consortia, a first requirement is thus a categorization of the jointly operated system.

1.1 Categorization of Cooperations

Terminology such as “digital preservation network”, “digital preservation collaborations” and “digital preservations services” have been used loosely, leading to no distinct boundaries between infrastructural and service levels associated with the terms. However, to fully understand the work conducted by a participating institution versus that being taken care of by a host or service provider, infrastructural and personal responsibilities need to be defined. Unfortunately no clear categorization schema exists as of today, leading to often misleading communication about networks, collaborations and jointly operated digital preservation systems.

The cost impact analysis put forth in section 2 of this paper uses the Curation Cost Exchange (CCEx) breakdown of digital preservation activities and resources. The author proposes to use this breakdown to further categorize jointly operated digital preservation systems, preservation networks and preservation services. To achieve this, the four CCEx service/activity categories Pre-Ingest, Ingest, Archival Storage, Access[3] – are used and further divided into the resource layers “Infrastructure” and “Preservation Management”. Infrastructure can be mapped to the CCEx “Cost by Resource” classification as containing purchases[4] and support/ operations staff (see Staff - Support Operations in Table 3). Similarly, Preservation Management can be mapped to the CCEx “Cost by Resource” classification as containing Producer and Preservation Analyst staff (see Staff - Producer. and Staff – Preservation Analyst in Table 3). To further exemplify: “Preservation Management” includes any human task associated with the digital object (as opposed to the preservation framework) along its lifecycle. This includes tasks such as defining packaging and mapping at the pre-ingest level, conducting deposits and handling errors occurring in file format validation steps at the ingest level, preservation planning and action at the archival storage level as well as defining DIPs (dissemination information packages) and access rules at the access level. Human tasks supporting the maintenance of the digital systems, such as system and network administration is captured on the infrastructural level.

[1] https://www.danrw.de/

[2] http://www.dpn.org/

[3] See Table 2

[4] see a)i), a)ii) and a)iii) in Table 3

The derived criteria are listed in the first column of Table 1. In a second step, each criterion is either assigned to the host level, meaning that the hosting or leading institution/ entity is responsible, or to the participating institution level. Table 3 shows a thus completed categorization view for the Goportis Digital Archive.

Table 1: Categorization of the Goportis Digital Archive. The criteria are based on the CCEx categories.

Criteria	Reponsibility
Pre-Ingest – Infrastructure	Participating institution
Pre-Ingest–Preservation Management	Participating institution
Ingest - Infrastructure	Host
Ingest – Preservation Management	Participating institution
Archival Storage – Infrastructure	Host
Archival Storage - Preservation Management	Participating institution
Access - Infrastructure	Host
Access – Preservation Management	Participating institution

1.2 The Goportis Digital Archive

TIB hosts the cooperatively operated digital preservation system for the Goportis consortium. The consortium consists of the three German national subject libraries: TIB Leibniz Information Centre for Science and Technology, ZB MED Leibniz Information Centre for Life Sciences and ZBW Leibniz Information Centre for Economics. Furthermore, TIB is currently designing a preservation-as-a-service offer for smaller institutions. The three Goportis partners finance the digital preservation system and the human resources responsible for it from their own resources, which are firmly fixed in each cooperation partner’s annual budget. The costs of jointly operating the system are currently borne equally by all three institutions. Each partner has its own digital preservation team that is firmly embedded in each institution’s structure and organisational chart. TIB is the Rosetta software licensee, hosts, operates and administers the digital preservation system, and provides Goportis partners access to the system. Use and operation are regulated in cooperative agreements between TIB, ZB MED and ZBW. [5]

Reflecting on the categorization put forth in Table 1, TIB covers both roles – participation institution, as the system is used for its own holdings, as well as host. It is important to stress that this paper is only written from the viewpoint of the host role.As the Goportis consortia falls into the smallest scale of networks, it is of utmost importance to check the impact which losing an institution would have on the network.

This paper puts forth first results of TIB’s analysis of risks associated with an institution leaving the consortia. The following sections highlight two key areas of risks: risks associated with the overall cost of the consortial operation of the Goportis Digital Archive and risks associated with the content belonging to the different institutions. The sections describe how the analysis was conducted and for both areas, cost and content, concrete risks are described including an impact evaluation as well as a first suggestion for mitigation strategies. While the sections 2 and 3 describe the analysis strictly from the viewpoint of TIB as the host of the consortial operation, the final conclusion and outlook section will touch on the relevance of this work to other institution and outline next steps which TIB intends to take.

2. COST RISKS

The last decade has seen a lot of research toward the cost of digital preservation [8]. While most institutions still show reluctance towards sharing cost information [9], various cost models have been put forth which allow institutions to evaluate their own cost requirements. For the evaluation of cost in the consortial context, the cost breakdown of the 4C project’s CCEx (Curation Cost Exchange)[6] platform was chosen as it is based on a gap analysis of prior cost model work done in other major projects such as LIFE³ and KRDS (Keeping Research Data Safe). CCEx allows the institutions to define a cost unit, and to allocate the total cost of that unit twice: once by service/activities and once by resources (purchases and staff) [9].

The breakdown for cost by service/activities can be taken from Table 2, which indicates the relevant criteria for TIB as the hosting institution (see also Table 2).

Table 2: CCEx Service/Activity levels and corresponding responsibility level of TIB as the hosting entity of the Goportis Digital Archive

	Service/Activity	Goportis Digital Archive responsibility
1.)	Pre-Ingest	none
2.)	Ingest	Infrastructure
3.)	Archival Storage	Infrastructure
4.)	Access	Infrastructure

Within the Goportis Digital Preservation System Pre-Ingest work is strictly done within the partnering institutions’ infrastructure. Data is transferred to the TIB environment for Ingest – relevant system architecture parts for the Ingest process are the network connection to the partnering institutions, allocated transfer storage as well as allocated operational storage which the digital preservation system requires for system internal ingest processes such as technical metadata generation. The archival storage is kept separate from the operational storage and keeps 2 copies plus backups. Automated processing mainly takes place during ingest and preservation action, including (re-)identification processes for file formats or the (re-)running of fixity checks. The system is currently operated as a dark archive and access only takes place for proof-of-concept purposes, for checks done by preservation

[5] See Goportis Digital Preservation Policy, available at: http://www.goportis.de/en/our-expertise/provision-of-scientific-content/digital-preservation/goportis-long-term-digital-archives-preservation-policy.html

[6] http://www.curationexchange.org/

staff or for trigger-based manual delivery of objects in case of corruption or loss of the access copy in use within external access systems. Due to this clear understanding of the resources currently used for the different activities, we can derive a rough estimate of cost percentage dedicated to the different services, as shown in Figure 1.

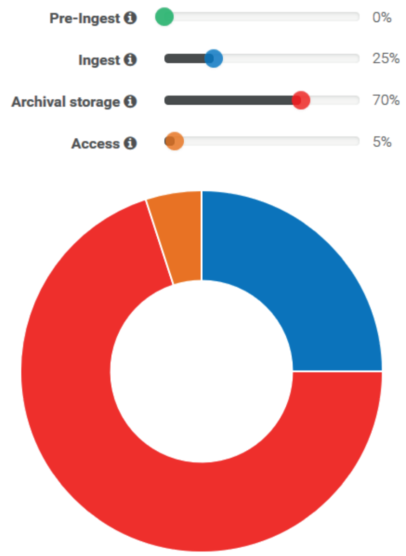


Figure 1. Estimate of cost breakdown by activity

The breakdown of cost by resources is hown in table 3. Here, the responsibility is matched to either to TIB as the host of the digital preservation system or to one or several of the participating institutions.

Table 3: CCEx Reource levels and corresponding responsibility level of TIB within the Goportis Digital Archive

Cost category	Cost	Responsibility
1.) Purchases	a) Hardware	Host
	a.) Software	Host
	b.) External or third party services	Shared by participating institutions
1.) Staff	a.) Producer	Participating institutions
	b.) IT developer	Participating institutions
	c.) Support/ operations	Host
	d.) Preservation Analyst	Participating institutions
	e.) Manager	Host
2.) Overhead	a.) Overhead	Host

While the hardware used has already been described in the analysis of "cost by service/activities", the software used is the proprietary digital preservation system "Rosetta" by Ex Libris for which the consortium shares the license cost. Further third party tools or services are currently not in use.

Within the digital preservation system the partnering institutions conduct the deposit, preservation planning and preservation action for their own content. Furthermore, each institution has full access to APIs[7] which allow the extension of the system to institutional needs. Development capacities within the institutions range between 0.25 and 1 FTEs (full-time equivalent). While developments may be used more than one institution, for example the development of a proxy mapping metadata imported from the union catalogue to the descriptive metadata, currently no dedicated consortial extension exists and the IT developer resource does not count towards the "consortial operation" cost unit. Support/operations, however, caters to all three partnering institutions. In addition to 1 FTE for system administration approx. 0.25 FTE go towards support of the partnering institutions for daily system operations including communication with the system vendor's support. Managerial work includes organizational coordination between the three institutions while overhead accounts for fixed costs such as office and server room space and electricity.

In addition to the cost unit break-down, CCEx requests a breakdown of the digital assets including an indication of type, size and volume [9]. As archival storage makes up a large cost factor, this analysis will be conducted per institution in the near future.

The break-down of the cost unit "consortial operation" by services/activities and resources allows for a good understanding of cost factors. Based on the high-level analysis, three cost risks can be determined, which are briefly discussed below: hardware/ infrastructure, software licenses and staff.

2.1 Hardware / Infrastructure

2.1.1 Risks

The estimate has shown that archival storage needs account for a large section of the overall costs. The requirements in archival storage size are naturally mandated by the archived content of the partnering institutions. In case of an institution leaving the consortium, the used storage space would be freed and would currently not be needed. The potential risk is that the infrastructure could be oversized for the existing requirements of a changing consortium constellation.

2.1.2 Impact

Impact depends on the overall size of the repository as well as the holdings and growth rates per institution. In the case of the Goportis digital preservation system the impact can currently be described as "low", as the freed storage can be easily allocated to the other two institutions without oversizing the repository or institutional storage allocation. Furthermore, TIB's infrastructure would allow free storage not used by the digital preservation system to be allocated to different services.

2.1.3 Mitigation Strategy

In addition to the CCEx recommended breakdown of digital assets in the as-is state, a prognosed growth rate per institution is collected on a yearly basis. It is advisable that the prognosis interval matches the notice period of the partnering institutions.

Furthermore, the break-down analysis of the cost-unit "consortial operation" shall be re-run once a year to check against new risks which can arise due to new requirements such as access to an institution's light archive collection.

2.2 Software Licenses

2.2.1 Risks

While a breakdown of purchase cost is currently not available, software vendor cost is always a key factor. The risk exists in form of license and support costs not tied to a specific number of institutions. In that case, an institution leaving the consortia would leave the remaining institution having to cover higher license and support costs.

2.2.2 Impact

Impact depends on the software license and support agreement, on the licensing and support cost as well as on the consortia size. As the Goportis consortium only consists of three institutions, the impact is defined as "high".

2.2.3 Mitigation Strategy

Include scenarios for changing consortia constellations and varying consortia sizes in the vendor contract.

2.3 Staff

2.3.1 Risks

The majority of staff for the consortial system goes towards system administration with additional requirements for support/operation and managerial tasks. The risk exists in form of staffing requirements being oversized when an institution leaves the consortia.

2.3.2 Impact

Impact depends on the overall size of the consortia and the staffing requirements based on that. In the case of the Goportis digital preservation system, support/operation as well as managerial tasks are covered by various TIB digital preservation team members who also perform institutional digital preservation tasks. The system administration FTE is required regardless of the size of the consortia. Due to this, the impact on staff can be described as "low".

2.3.3 Mitigation Strategy

Staffing requirements for consortial operation shall be re-evaluated on a yearly basis to check for changing risks. Spreading out support/operation and managerial tasks across different staff minimizes the risk of an oversized team structure.

3. CONTENT RISKS

An institution leaving a consortia is a concrete exit scenario. A solid exit strategy is an integral part of every digital preservation system. Certification processes such as TRAC [10], the Data Seal of Approval [11] and the nestor seal [12] require or recommend that exit strategies be in place. However, certification guidelines do not give concrete description of what exit strategies should contain. Instead, the strategy is usually considered evidence of appropriate succession and contingency plans. Commonly, the use of systems which support open standards is seen as a pre-requisite for an exit strategy [1]. However, current descriptions of exit scenarios usually pertain to the situation where an existing institutions exits from one system into another. Contingency plans covering the institution's demise usually only focus on technical requirements for data export, such as completeness and open formats, as well as extensive representation information to allow for adequate interpretation of the digital objects. Legal aspects are highly specific to the jurisdiction of the archive and are less frequently covered in exit strategies [13][1].

As opposed to a system-wide exit scenario, a consortially operated system calls for a tiered exit scenario which clearly allows for the export and interpretation of the data pertaining to a single institution. Furthermore, two scenarios need to be considered: the institution exits because it leaves the consortia but continues to exist and the institution exits because it ceases to exist. In the latter case, the data may need to be handed over to a third-party which leads to different legal requirements and implications.

These legal implications as well as standard exit scenario requirements lead to four risks associated with the content of an institution leaving a consortium. These risks are further described in the following subsections.

3.1 Export of Institutional Data

3.1.1 Risks

In the case of an institution exiting a consortium the repository needs to be able to export and delete the institution's data from the repository while leaving the data of the remaining institutions intact. The risk is that the repository is either unable to select the objects and their associated metadata per institution and/or that the exported data is incomplete or not interpretable outside of the digital preservation system.

3.1.2 Impact

This risk exists for any consortium, regardless of size or makeup. As the repository operator would not be able to fulfill a fundamental requirement of a trustworthy digital preservation system the impact has to be defined as "high".

3.1.3 Mitigation Strategy

A consortial system shall clearly differentiate between the different institutions from the start. Ideally, different data management interfaces exist for the different institutions. Workflows shall be completely separated and the objects' accompanying metadata shall clearly include the institution as the content owner. Additionally, separate storage locations should be set up for each institution.

3.2 Documentation of Institutional Processes

3.2.1 Risks

Preservation processes may include documentation which is not directly stored within the repository. Examples for this are full license agreements between a depositor and the institution. While the license text may be included in rights metadata, the signed agreement is usually stored in a rights management system or resides as a hard-copy within the institution. Another example is supporting documentation for a preservation plan.

While not directly available within the repository, this information is still essential for interpretation of the digital objects across their lifecycle. Especially in the case where an institution exits the consortium due to its demise and the digital objects are to be handed over to a new steward, either a consistent link to external information or, ideally, the entire information itself, shall be provided in a data export.

3.2.2 Impact

The impact is especially "high" for the archiving institution as well as for a potential third party who takes over as a steward of data in the case of the institution's demise.

3.2.3 Mitigation Strategy

Consortia wide policies shall be in place to regulate the availability of complementary information for all preservation workflows. Where it is not possible to store the information in the repository, a clear description of where to find the information must be given.

3.3 Non-transferable Rights

3.3.1 Risks

No risk exists if an institution exits and requests an export of their objects to store in a different system or locally. However, the situation is different if an institution exists because it ceases to exit. In that case, a new steward for the institution's objects needs to be found and the consortium leader may therefore have to pass the objects on to a third-party. The risk here resides in often non-transferable rights of digital objects [14].

3.3.2 Impact

The impact is particularly "high" for a future steward of information which previously belonged to an institution which ceased to exist. Unless the objects are licensed under a public

[7] https://developers.exlibrisgroup.com/rosetta/apis

license, the license will have to be re-negotiated between the creator and the data steward. This becomes particularly hard if the information provided about the creator alongside the object is only rudimentary.

3.3.3 Mitigation Strategy

While there is no solution for non-transferable rights, the situation can be improved by including further information about the creator. Here, particularly contact information such an email address is helpful. Also, the availability of the full original license agreement, as described in section 3.2, is beneficial.

3.4 User Names in Metadata

3.4.1 Risks

As part of PREMIS based preservation metadata generation, the Goportis Digital Archive gathers information about agents. These agents can be software as well as users. If a user acts as an agent, the username is captured in the metadata. If a user performs a deposit, additional information such as the full name, work address and email are captured. Full address information of the user is also included in the user's profile.

In Germany the use of personal data is protected by the BDSG (Bundesdatenschutzgesetz) law book. BDSG §20 states that public institutions – such as the three Leibniz information centres belonging to the Goportis consortia – are required to delete personal data of their employees as soon as this data is no longer required to fulfill its original purpose [15]. As in the case of non-transferable rights this becomes especially a problem when an institution exits due to its demise and the objects and their accompanying metadata are to be handed over to a third-party as the new data steward. Since the preservation metadata is an integral part of the AIP to be handed over, all user data captured within would need to be anonymized or pseudonymized.

3.4.2 Impact

As described above, the impact is "high" if the objects need to be handed to a third party who becomes the new data-steward.

3.4.3 Mitigation Strategy

An overview of where user data is captured within the metadata shall be prepared to assist in an anonymization process. It needs to be evaluated if pseudonymization is preferable, e.g. by substituting user names by a fixed set of roles. The understanding of what role triggered an event within a workflow may assist a third-party institution in better interpreting the preservation metadata as well as the lifecycle events it describes.

4. CONCLUSION AND OUTLOOK

While the analysis of the impact which an institution leaving the consortium imposes is still a work-in-progress, this paper put forth a first analysis of risks associated with the overall costing of the cooperatively operated digital archive as well as of risks associated with the content of the institution exiting.

In regards to the cost analysis, the CCEx tool proved to be extremely helpful in analyzing affected cost segments. Here, further work will be invested in two tasks: (a) gather information to allow for a better differentiation between economic and non-economic cost factors[8] and (b) a detailed analysis of the holdings per size, type and volume for each institution including effective growth over the past two years and prognosed growth for the next year

Regarding the content analysis, the results made clear that the extent of on object's description in its lifecycle – especially when the lifecycle shall foresee a transfer to a different data steward – are wider than anticipated. The two take-aways here are: (a) the Goportis digital preservation policy should be checked towards including further information regarding the availability of relevant object lifecycle information currently not stored in the repository and (b) the export of all institutionally relevant data shall be checked regularly including a strategy to anonymize or pseudonymize the user data captured in the preservation metadata.

Also, further work will go into the identification of other impact areas. The impact on "shared knowledge and efforts" is one which is currently not yet covered. For example, the Goportis Digital Archive shares networking activities and maintains a wiki to exchange results. Losing a partner would impact this form of knowledge aggregation.

The analysis in this paper was strictly conducted from the viewpoint of TIB in its role as the consortial leader and host of the Goportis Digital Archive. As such, the situation evaluated was that of TIB losing a partnering institution. Needless to be said the situation would be completely different if the institutions would lose their consortial leader and host. Despite the specific use case given here in form of a small network of three large national subject libraries, the identified risks shall apply to preservation collaboration or networks of different make-up and size. An analysis of the cost unit "consortial operation" for a different network will most likely lead to different distribution results regarding service/activities and resources as other networks may very well include pre-ingest work or share IT development resources. However, the risk breakdown of "hardware", "software" and "staff" appears to be a universal one and while the impact may of course differ, the briefly sketched mitigation strategies may be used a basis for own work. The impacts of the content and the associated risks seem to be universal regardless of preservation makeup and size. While legislation differs from country to country, the transferability of rights and the requirements to anonymize user data should still be checked.

5. REFERENCES

[1] Corrado, E., Moulaison, H.L. 2014. *Digital Preservation for Libraries, Archives, and Museums*. Rowman & Littlefield., Lanham, MD.

[2] Dappert, A. *Risk Assessment of Digital Holdings*. TIMBUS Project presentation. http://timbusproject.net/documents/presentations/9-risk-management-and-digital-preservation

[3] Graf, R., Gordea, S. 2013. A Risk Analysis of File Formats for Preservation Planning. In *Proceedings of the 10th International Conference on Preservation of Digital Objects*. IPRES2013. Lisbon, Portugal.

[4] Graf, R., Gordea, S. 2014. A Model for Format Endangerment Analysis using Fuzzy Logic. In iPRES 2014 – *Proceedings of the 11th International Conference on Preservation of Digital Objects*. iPRES 2014. Melbourne, Australia.

[5] Innocenti, P., McHugh, A., Ross, S.. 2009. Tackling the risk challenge: DRAMBORA (Digital Repository Audit Method Based on Risk Assessment), In *Collaboration and the Knowledge Economy: Issues, Applications, Case Studies*. Cunningham, P., Cunningham, M. (Editors). Stockholm, Sweden.

[8] EU legislature requires publically funded institutions to clearly separate economic and non-economic activities in financial reporting. Non-profit entities need to have a detailed auditing for all processes going towards services such as hosting.

[6] Vermaaten, S., Lavoie, B., Caplan, P. 2012. Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment. *D-Lib Magazine*. Volume 18, Number 9/10 (September/October 2012).

[7] Archives New Zealand. 2010. *Government Digital Archive – "Rosetta" Gap Analysis – Update*. Version 2.0.Technical Report. 18th August 2010.

[8] 4C Project. 2013. *D3.1 – Summary of Cost Models*. Technical Report. 4C Project.

[9] Middleton, S. 2015. *D2.8 – Curation Costs Exchange*. Technical Report. 4C Project.

[10] CRL/OCLC. 2007. *Trustworthy Repositories Audit & Certification: Criteria & Checklist*. Technical Report. CRL/OCLC. Chicago, IL and Dublin, OH.

[11] Data Seal of Approval Board. 2013. *Data Seal of Approval Guidelines version 2*. Technical Report.

[12] nestor Certification working group.2013. *Explanatory notes on the nestor Seal for Trustworthy Digital Archives*. Nestor-materials 17. nestor.

[13] Schaffer, H. Will You Ever Need an Exit Strategy? In *IT Pro*. 4-6. March/April 2014.

[14] Euler, E. 2011. *Digitale Bestandserhaltung und Distributed Storage in LukII*. Legal report. DFG

[15] Federal Republic of Germany. *Bundesdatenschutzgesetz (BDSG)*. In der Fassung der Bekanntmachung vom 14.01.2003 (BGBl. I S. 66). Zuletzt geändert durch Gesetz vom 25.02.2015 (BGBl. I S. 162) m.W.v. 01.01.2016

A Method for Acquisition and Preservation of Emails

Claus Jensen
The Royal Library
Søren Kierkegaards Plads 1
Copenhagen, Denmark
+45 91324448
cjen@kb.dk

Christen Hedegaard
The Royal Library
Søren Kierkegaards Plads 1
Copenhagen, Denmark
+45 91324525
chh@kb.dk

ABSTRACT

In this paper we describe new methods for the acquisition of emails from a broad range of people and organisations not directly connected with the organization responsible for the acquisition. Existing methods for acquisition of emails are based either on having easy access to an institution's email server, or a labour intensive process of transferring the emails from donors' email clients, involving for example visits to the individual donors.

Furthermore, we describe how different representations of the acquisitioned emails are ingested into our repository. The use of different representations makes it possible for us to perform a fine grained file characterisation, thereby facilitating the level of preservation watch we want in connection with the preservation of the acquisitioned emails.

Keywords

Email; Acquisition; Preservation; Repository: Linked data.

1. INTRODUCTION

The email project began with a request from the curators in The Royal Library Manuscript Collections to acquire emails from individuals in arts and sciences (scholars, authors, etc.). The request was based on the assumption that much of today's written memory is born digital and that there is a high risk of losing this material if the acquisition process does not start until after the donor has passed away. Some of the major threats to the material are deletion due to computer crash, and limited space on email servers.

The primary audience for the final service is Danish archive creators such as authors, researchers, artists, and persons and organizations active in the cultural domain taken in a broad sense. Their digital material is considered important for future research in particular in areas of science, the history of learning, cultural history, etc.

The curators in The Royal Library Manuscript Collections have analyzed their audience to fall within three major groups: The first group are employees in institutions which almost exclusively use their institutions' email system. The second group are also employees of institutions, but this group mostly use their own private email. The third group is not affiliated with an institution, and therefore only use their private email.

As most of the target group was in the latter two groups, it was not possible to use the acquisition method where access goes through an institution's email system. The method of acquiring email from the individual donors' email clients was considered far from optimal both from a labour resource perspective and from a technical perspective.

2. STATE OF THE ART

A survey of methods for acquisition and preservation of emails can be found in the DPC Technology Watch Report Preserving Email [7]. A series of articles concerning the acquisition and preservation of emails has been written [1], [2], [3], [4], [9], [10], [11], [12]. The articles do not always describe the exact method of acquisition, i.e. how the emails are transferred from the donors to the institution responsible for the acquisition. However, even when the method is not explicitly described, it is often possible implicitly to see what methods have been used. The two most widely used methods of acquisition are: To extract the emails from email servers from which the institution has easy access or to use a more labour intensive process involving acquisition of emails through the donors' email client.

Different methods on how to pre-process and ingest emails into a repository have been studied in a number of articles. In E-mails to an Editor [1] it is described how the project ingest emails into a repository in three different formats MSG, EML, and XML and the Aid4Mail program [6] is used for the pre-processing of the emails. In Reshaping the repository [2] the process of how the project converts emails into the RFC-282 Internet Message Format [8] using the Emailchemy program [13] is described. In Coming to TERM [3] it is described how emails are converted to the RFC-282 Internet Message Format, if the original format is a proprietary format. The email and its attachments are marked up in XML before they are ingested into the repository.

3. THE INITIAL REQUIREMENTS

The process of finding or building a system for the acquisition of emails was initiated by a phase of collecting requirements with input from both the curators and us. The curators had a series of mostly non-technical requirements for the new email acquisition system.

Table 1. Non-technical requirements

Maximum emulation of the traditional paper-based archiving criteria and procedures
High level of security against loss, degradation, falsification, and unauthorized access
A library record should exist, even if documents are not publicly available
Simple procedure for giving access to third-party by donor
Maximum degree of auto-archiving
Minimum degree of curator interference / involvement after agreement

Similarly, we had a number of technical requirements for the system.

Table 2. Technical-oriented requirements

No new software programs for the donor to learn
No installation of software on the donor's machine, and if programs had to be installed, it should be standard programs and not programs we would have to maintain
As much control over the complete system in our hands as possible
As much as possible of the workflows within the system should be automated
Independence from security restrictions on the donor system imposed by others (password secrecy, restrictions on installation of programs, etc.)

4. THE FIRST PROTOTYPE

The first prototype was implemented on The Royal Library email system for a limited number of donors, selected by the curators. Each donor was given an "archiving email account".

We allowed the donors to choose between different methods for archiving emails. One of the methods was adding their archiving account as a BCC recipient when sending or responding to an email. Another method was to forward received or sent emails to the archiving account. The use of forwarding would for example be necessary when donating the last received email in a thread.

The donors chose to employ two different processes: One group of donors donated their emails using a continuous process of sending and receiving emails by using BCC and forwarding. The other group used a periodic donation process. An example of the use of the periodic process was when donors donated on a monthly basis by forwarding the emails to their archiving account.

A major disadvantage of the forward method for archiving emails is that important information contained in the original email header is either lost or hidden inside unstructured email message text. For the curators the original date of the email was important.

In some cases it would be possible to extract the send date of the email from the email message, as a number of email clients use a semi-structured way of registering this information within the email message. However, the email clients used different methods to separate the send-date information from the rest of the email message. Therefore it was not possible to implement a general method to extract the original send-date information.

Other disadvantages of using the forward method for archiving emails that we encountered were:

- It was easy for the donor to forget to forward the last message in an email thread
- Periodical donation sometimes failed because the email "package" was too big due to the following reasons:
    - ○ A timeout from the antivirus scanner because the scanning time of the email exceeded the maximum time period allowed
    - ○ The email provider had a size limit on the emails

We had to conclude that the first prototype had some serious drawbacks. Thus we had to look for other solutions for the acquisition of the donors' emails.

5. THE SECOND PROTOTYPE

Using our experiences from the first prototype and combining them with new ideas for the acquisition process, a new series of requirements took form in the beginning of the second phase of the project. In formulating the new requirements, we drew on both the donor's and our own experiences with the first prototype.

The additional requirements were formulated in the following way.

Table 3. New requirements for the second prototype

The system should be based on standard email components
Easy to use for both curator and donors
No curators should have to visit the donors' residence for setup or email transfer (self-deposit)
The system should be based on voluntary/transparent deposit
It should be independent of technical platforms (PC, Mac, iOS and Android devices, etc.)
The donor should have the option to transfer emails to the deposit area at any time
The donor should always have access to their donated emails
Based on permission granted by the donor different levels of access for external use should be allowed at any time.
The donors must be able to organize and reorganize emails.
The donors must be allowed to delete emails in the system within a certain time-frame
The original email header metadata must be preserved
The donors must be able to deposit other digital materials along with their emails

During the new requirement process it became increasingly clear that it was necessary to create two areas for each donor. We named these areas, respectively, the *deposit area* and the *donation area*. The deposit area was defined as a temporary dynamic email area where the donor (also called the "archive creator") could transfer all of their emails from their different email accounts. Furthermore, the archive creator still has all rights to the materials in their deposit area and is able to edit the deposited emails (create new emails and folders, move emails and folders, delete emails and folders, copy emails and folders, etc.).

The desired time period for the deposit of emails is specified in the agreement between the curator and the donor. Typically a three year deposit period is chosen. When the archive creator is ready to donate, the curator moves the agreed emails from the deposit area to the donation area. The emails then become the property of The Royal Library. The donor (previously archive creator) will now only have read access to their emails. After this the donated emails are ready for ingest into the repository system as part of the long-term preservation process.

The new requirements initiated a major redesign of the system. We decided to continue the principle that every donor should have their own email account. The open question on how to transfer the donors' emails to their archiving accounts without losing important information remained.

We investigated the possibility of using the email clients' ability to handle more than one email account at a time. This ability does not only mean that it is possible to read and write emails in connection with many email accounts, but also support the process of moving and copying emails and folders between different email accounts. The moving or copying of emails from one email account to another within the email client itself does a much better job of preserving the important information we lost in the first prototype.

To support as many email clients as possible we decided to use the IMAP (Internet Message Access Protocol) and SMTP (Simple Mail Transfer Protocol) between email clients and email servers. The IMAP protocol is implemented in all widely used email servers and email clients and it is platform independent. Furthermore, the IMAP protocol is both supported by the email clients of modern smart devices and by the many free email clients for computers. Even though it is not possible to transfer emails directly from web-based email systems such as Gmail and Yahoo Mail, it is possible to transfer these emails using an email client supporting the IMAP protocol.

The process of moving and copying email and creating new folders within a single email account are well-known tasks for most donors. Therefore it was expected that these processes would be easy to perform for the donors even though they now had to perform these tasks between two email accounts instead of only a single account.

The second prototype allows the donor group that prefers a continuous donation process the ability to copy and paste (drag and drop) single emails to their archiving account immediately after they either send or receive new emails. The other group of donors who prefer using a more periodic donation process would have the ability to copy and paste multiple files or folders to their archiving account using larger time intervals.

Our second prototype was implemented as an independent email server, in our case an Exchange server [14], totally separated from The Royal Librarys' email system. Furthermore, the deposit area was separated from the donation area

The service options of the Exchange email server were limited as much as possible. Available service options were

- Full access via IMAP
- Webmail, but limited to read access and for changing the password for the email account.

The email accounts were set up so they could not receive emails. This was done to avoid unauthorized email messages like spam emails getting into the deposit area.

The new method of acquisition gave the donors the following benefits:

- They were able to use their own email client (Outlook, iOS mail, Thunderbird, etc.)
- They could deposit via different devices (Windows, Linux, iOS devices, Android devices, etc.)
- They could use several devices for depositing emails.

There were now only the following requirements for donors to deposit their emails:

- The donor must have access to an email client
- They must be able to setup an IMAP account in their email client on their own device.

The configuration of the IMAP and SMTP connections was, due to internal IT-polices at our institution, non-standard. The non-standard configuration resulted in the need to use a more complicated configuration for most of the used email clients. However, the latest developments in modern email clients has resulted in, that much of the complicated configuration can be done in an automated way, where only basic information like email address, user name, and email-server name need to be inserted by the user.

6. FROM DEPOSIT TO DONATION

At a given time (based on the agreement between the donor and the receiving institution) the deposited material becomes the property of the institution and is transferred to the donation area. In our setup the donation area is another email server where the curators can work with the donated emails. This means that the curators can process the emails in a familiar environment using the tools they normally use for handling their own emails. When the curators have finished processing the donated emails, the complete email account is exported to an email account file container (we currently use the PST file format) and this file is then ready for further processing and ingest into our repository system.

7. EXPERIENCES WITH THE SECOND PROTOTYPE

The experiences with the second prototype, which has become the current production system, were much better for everyone involved: donors, curators, and system managers. The curators could work with the donated emails in the same way that they work with their own email, and the work process was easy and well-know. Similarly the donors had the same experience in their donation process which they also found easy and familiar.

The configuration of their email account on their own devices caused problems for many donors. Even though the configuration of the email account only had to be carried out once, we had to put a lot of effort into the user manual. This part of the system was not completely standard as we for security reasons was using other ports and encryptions than the ones most email clients employ as defaults.

Many of the donors did not want to read the user manual, particularly when it came to setting up port numbers and encryption standards. Furthermore, given the many different email clients in different versions it was not possible to write documentation for every single one, and this complicated the configuration process for some donors.

In most cases the curators were able to help the donors with the email-client configuration. When a donor's email client was properly set up, no further problems were observed in the depositing process itself.

The new method of depositing emails provided a more intuitive way of depositing for those donors who prefer a periodical process. At the same time the difficulty of depositing emails for the donors who prefer a continuous deposition process was not increased when comparing with the first prototype where depositing was done using BCC or forward.

Furthermore, the new method of depositing emails has the advantage that the donor can easily organize their emails into folders or upload entire folders if they prefer. In addition to this the donor has full access to the email account and can also delete emails if they want.

8. INGESTING EMAILS INTO OUR REPOSITORY

When we ingest the received emails into our repository, we employ some of the same tools used by institutions having similar ingest workflows, e.g. The University of Manchester Library [1]. However, the way we use these tools and particularly the way our repository is structured is very different. We ingest the donated emails into our repository system (which is based on Hydra [15] and Fedora Commons [16] version 4). Different representations of the email account are ingested. The email container file is one representation and this representation is ingested manually by our curators using the repository's web interface for upload of files and addition of metadata. We also ingest another representation of the email account where the account has been "unfolded" into its parts (folders, emails, attachments, and their relations). See the sketch in Figure 1 for an example case. The transformation from the container representation to the multi-parted representation is done using the Aid4Mail program [6]. A specialized script has been produced that bundle the different Aid4Mail processes and extract additional metadata.

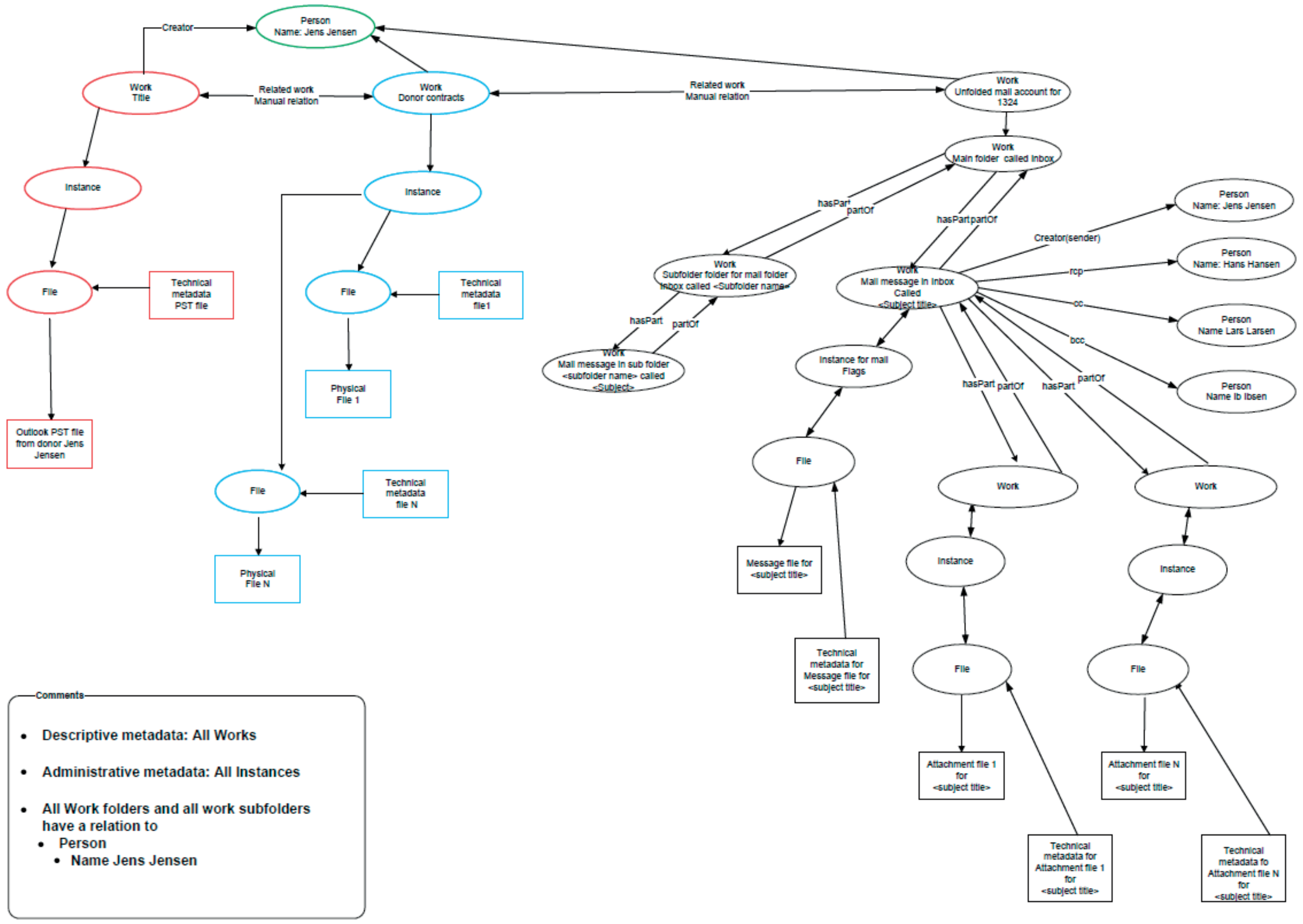

Person
Name: Jens Jensen
Creator
Work
Title
Related work
Manual relation
Work
Donor contracts
Related work
Manual relation
Work
Unfolded mail account for 1324
Instance
Instance
File
Technical metadata PST file
Outlook PST file from donor Jens Jensen
File
Technical metadata file1
Physical File 1
File
Technical metadata file N
Physical File N
Work
Main folder called Inbox
hasPart
partOf
hasPart partOf
Work
Subfolder folder for mail folder Inbox called <Subfolder name>
hasPart
partOf
Work
Mail message in sub folder <subfolder name> called <Subject>
Work
Mail message in Inbox Called <Subject title>
Creator(sender)
rcp
cc
bcc
Person
Name: Jens Jensen
Person
Name: Hans Hansen
Person
Name Lars Larsen
Person
Name Ib Ibsen
Instance for mail Flags
File
Message file for <subject title>
Technical metadata for Message file for <subject title>
hasPart partOf
hasPart partOf
Work
Work
Instance
Instance
File
File
Attachment file 1 for <subject title>
Attachment file N for <subject title>
Technical metadata for Attachment file 1 for <subject title>
Technical metadata fo Attachment file N for <subject title>
Comments
Descriptive metadata: All Works
Administrative metadata: All Instances
All Work folders and all work subfolders have a relation to
Person
Name Jens Jensen

Figure 1. Handling of email objects and metadata

Another product of the transformation is a XML representation of the email container which contains structural information, email-header information, the email body in Unicode text format, and information about the individual emails attachments. We use this XML representation to generate metadata for the individual objects (folders, emails, and attachments) and their relations when ingesting them into our repository.

9. LINKED DATA AND EMAILS

Our repository supports linked data and uses RDF within its data model. We use this feature to create relations between the objects. For example: hasPart and its reverse partOf holds the relationship between folders and emails and between emails and attachments. Furthermore, we use RDF relations to connect emails with agents where the agents act as sender or recipient of the emails.

In the long-term perspective this use of linked data can connect not only our donors internally within our system, but in principle also our donors to other email collections in other institutions. This means that networks with the correspondence of, for example, a group of researchers can be formed.

In a preservation context ingesting the different email representations into our repository system provides the possibility to perform file characterisation on all the parts of the email collection; the email container files, individual emails, and attachments. The ability to do this characterisation on the whole content allows us to perform preservation watch. If we only ingested the container file we would not be able to perform a complete preservation as currently no characterization tools are able to unpack the container file and perform a characterization on its individual objects. The cost of this approach is obviously an increase in the amount of storage (roughly doubling it). However, we can still decide not to long-term preserve every representation so there is not necessarily an increase in the storage cost for long-term preservation.

Having a multi-parted representation in our repository also allows us to preserve individual emails or attachments, or groups of these, at different preservation levels. The different preservation levels could for example consist of giving a particular selection of objects a higher bit safety. Furthermore, in a dissemination context where there are restrictions on the email container, the restrictions on individual emails or attachments or groups of these can be lowered, making it possible to disseminate them to a much broader audience.

10. FUTURE WORK

The email project is still active, and there is still time to explore alternative or supplementing methods for the acquisition of emails. Also the task of finding good ways of disseminating the email collections has not yet begun.

10.1 Alternative Acquisition Methods

An alternative or supplementary way of acquiring our donors' emails could be to harvest them. This could be done in a similar way to the one we employ in our web harvests. This process would require the use of the IMAP protocol and therefore the use of other tools than the ones used in a standard web harvesting would be necessary. Challenges concerning authentication in connection with harvesting of a donors' email account would also have to be solved. A simple proof of concept has been made and the method is worthy of further investigation.

We are also interested in allowing the deposit of other digital materials. These could be video and audio files which in general are large in size. Even though the IMAP protocol supports transfer of large (in size) attachments, our experience is that it is not the best protocol for the task, as the performance in general is poor.

Instead a possibility could be to use a “Dropbox like” solution; another could be the use of sneakernet (physically moving media like external hard drives or similar devices).

10.2 Dissemination of Emails

At the current phase in the project we have only just begun considering the possibilities for a dissemination of the acquired emails. We considering two tools for this purpose: a standard email client (like Outlook [17]) and ePadd (formerly known as MUSE) [4], [5], [18].

The use of Outlook or similar email clients will give the end-user a well-know experience in which the search and reading of emails would be done in the same way as when the user handles their own email. The use of ePadd gives a greater series of possibilities for the users such as entity extraction, easy browsing and thematic searching. However with new possibilities come new features to be learned by the user, so this option would most likely mean more work both for the users and the curators.

Other alternatives or supplements to these tools should also be considered and tested, but our starting point will be the testing of the two above mentioned tools in collaboration with our curators and users.

11. ACKNOWLEDGMENTS

We would like to thank all that have helped with the development of the service, especially the members of the email project team.

12. REFERENCES

1. Fran Baker. 2015. E-mails to an Editor: Safeguarding the Literary Correspondence of the Twenty-First Century at The University of Manchester Library. *New Review of Academic Librarianship* 21, 2: 216–224. http://doi.org/10.1080/13614533.2015.1040925
2. Andrea Goethals and Wendy Gogel. 2010. Reshaping the repository: The challenge of email archiving. *iPRES 2010*: 71.
3. Marlan Green, Sue Soy, Stan Gunn, and Patricia Galloway. 2002. Coming to TERM: Designing the Texas Email Repository Model. *D-Lib Magazine* 8, 9.
4. Sudheendra Hangal, Peter Chan, Monica S. Lam, and Jeffrey Heer. 2012. Processing email archives in special collections. *Digital Humanities*.
5. Sudheendra Hangal, Monica S. Lam, and Jeffrey Heer. 2011. MUSE: Reviving Memories Using Email Archives. *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, ACM, 75–84. http://doi.org/10.1145/2047196.2047206
6. Fookes Software Ltd. Aid4Mail : Reliable Email Migration, Conversion, Forensics and More. Retrieved April 12, 2016 from http://www.aid4mail.com/
7. Christopher J. Prom. 2011. *Preserving email*. Digital Preservation Coalition.
8. P. Resnick. 2001. *RFC-2822 Internet Message Format*. The Internet Society.
9. B. Srinija, R. Lakshmi Tulasi, and Joseph Ramesh. 2012. EMAIL ARCHIVING WITH EFFECTIVE USAGE OF STORAGE SPACE. *International Journal of Emerging Technology and Advanced Engineering* 2, 10.
10. Arvind Srinivasan and Gaurav Baone. 2008. Classification Challenges in Email Archiving. In *Rough Sets and Current Trends in Computing*, Chien-Chung Chan, Jerzy W. Grzymala-Busse and Wojciech P. Ziarko (eds.). Springer Berlin Heidelberg, 508–519.
11. Frank Wagner, Kathleen Krebs, Cataldo Mega, Bernhard Mitschang, and Norbert Ritter. 2008. Email Archiving and Discovery as a Service. In *Intelligent Distributed Computing, Systems and Applications*, Costin Badica, Giuseppe Mangioni, Vincenza Carchiolo and Dumitru Dan Burdescu (eds.). Springer Berlin Heidelberg, 197–206.
12. Frank Wagner, Kathleen Krebs, Cataldo Mega, Bernhard Mitschang, and Norbert Ritter. 2008. Towards the Design of a Scalable Email Archiving and Discovery Solution. In *Advances in Databases and Information Systems*, Paolo Atzeni, Albertas Caplinskas and Hannu Jaakkola (eds.). Springer Berlin Heidelberg, 305–320.
13. Emailchemy - Convert, Export, Import, Migrate, Manage and Archive all your Email. Retrieved April 19, 2016 from http://www.weirdkid.com/products/emailchemy/
14. Secure Enterprise Email Solutions for Business | Exchange. Retrieved April 13, 2016 from https://products.office.com/en-us/exchange
15. Hydra Project. *Hydra Project*. Retrieved February 23, 2016 from http://projecthydra.org/
16. Fedora Repository | Fedora is a general-purpose, open-source digital object repository system. Retrieved February 23, 2016 from http://fedoracommons.org/
17. Email and Calendar Software | Microsoft Outlook. Retrieved April 5, 2016 from https://products.office.com/en-US/outlook/email-and-calendar-software-microsoft-outlook?omkt=en-US
18. ePADD | Stanford University Libraries. Retrieved April 5, 2016 from https://library.stanford.edu/projects/epadd

Processing Capstone Email Using Predictive Coding

Brent West
University of Illinois
506 S Wright St, M/C 359
Urbana, IL 61801 USA
+1 217-265-9190
bmwest@uillinois.edu

Joanne Kaczmarek
University of Illinois
506 S Wright St, M/C 359
Urbana, IL 61801 USA
+1 217-333-6834
jkaczmar@illinois.edu

ABSTRACT

Email provides a rich history of an organization yet poses unique challenges to archivists. It is difficult to acquire and process, due to sensitive contents and diverse topics and formats, which inhibits access and research. We plan to leverage predictive coding used by the legal community to identify and prioritize sensitive content for review and redaction while generating descriptive metadata of themes and trends. This will empower records creators, archivists, and researchers to better understand, synthesize, protect, and preserve email collections. Early findings and information on collaborative efforts are shared.

Keywords

Archives; Continuous active learning; Dataless classification; Descriptive metadata; E-discovery; FOIA; Metrics-based reappraisal; MPLP; Natural language processing; Restricted records; Self-appraisal; Sustainable digital preservation; Technology-assisted review.

1. INTRODUCTION

The Records and Information Management Services (RIMS) office of the University of Illinois is leading a project to help archivists preserve email messages of enduring value, beginning with those of the University’s senior administrators [1]. Email messages of senior administrators are the modern equivalent of correspondence files, long held to have enduring value for administrators and researchers alike. However, email presents unique accessioning challenges due to its quantity, file formats, conversation threads, inconsistent filing, links and attachments, mix of personal and official communications, and exposure of sensitive content.

The quantity and mix of content, as well as the inability to rely upon administrators to consistently identify messages of enduring value, led RIMS to explore the Capstone approach developed by the United States National Archives and Records Administration (NARA) [2] to stem the loss of significant correspondence. The Capstone approach offers an option for agencies to capture most of the email from the accounts of officials at or near the head of an agency without detailed consideration of the content.

Although this approach can help to ensure that significant correspondence is retained, Capstone is just the first step in the overall curation lifecycle [3] at a scale which necessitates More Product, Less Process [4]. Processes and tools such as Preservica exist to acquire, ingest, store, transform, and even provide access to email. However, unmet lifecycle challenges of email include the identification of restricted records as a prerequisite to public access and the reappraisal of non-archival messages in heterogeneous email collections. Techniques such as Metrics-Based Reappraisal [5] can sustainably inform reappraisal decisions for a variety of digital collections. However, we propose a new methodology to address both unmet challenges.

2. PREDICTIVE CODING

2.1 E-discovery

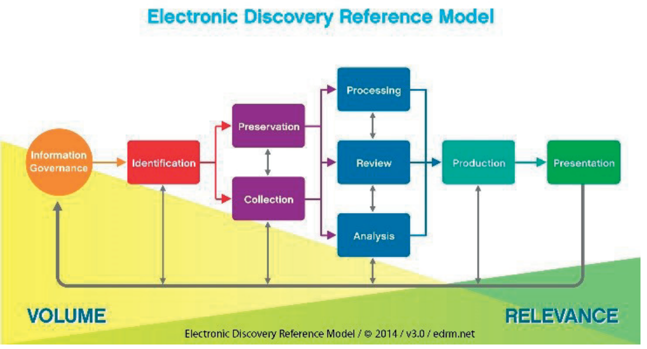


Figure 1. Electronic discovery reference model. [6]

Electronic discovery is a “process in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in a civil or criminal legal case” [7]. Courts require good faith efforts to discover and produce relevant evidence for the opposing party to a lawsuit. E-discovery provides attorneys insight into both their case and their opponents’ case, uncovering critical evidence that can resolve the case in one’s favor. With potentially millions of dollars on the line, the legal community has a substantial incentive to conduct a thorough review. At the same time, courts recognize that the burden of discovery must be proportional to the potential evidentiary value, the amount in dispute, and the resources of the parties. Even so, e-discovery is expensive with mean costs comprising 73% or $22,480 per gigabyte reviewed [8]. To combat these high costs and provide a competitive advantage, attorneys and courts are increasingly turning to technology to make the review process more efficient.

2.2 Technology-Assisted Review

Technology-assisted review (TAR) enhances the heretofore manual review of potentially relevant records by providing insight into data collections. TAR allows attorneys to more quickly locate potentially responsive documents and cull that list based on various attributes to narrow and prioritize review and redaction efforts. TAR tools often feature de-duplication, email threading, full-text search of messages and common attachments, and pattern and trend visualizations. Increasingly, TAR tools are providing natural language processing and machine learning features to cluster documents by topics and identify hidden relationships through predictive coding.

2.3 Predictive Coding

Predictive coding leverages artificial intelligence algorithms to locate relevant documents. Relevant documents that have been assessed manually by humans are processed by the algorithms to automatically assess the relevance of other documents in a large collection. In an iterative process of automated assessment and review, the software begins to learn what attributes make a document relevant, increasing the capacity to quickly identify documents of most interest. A variation of this approach is known as Continuous Active Learning [9] where the process is repeated until no further items shown are relevant. This ability to automatically categorize hundreds of thousands to millions of documents greatly enhances the effectiveness of document review, allowing attorneys to prioritize their review around the most valuable or sensitive content.

In a sense, predictive coding is automating the generation of topical descriptive metadata. The identification of documents that are relevant to particular topics allows archivists to prioritize the review of a large email collection and identify restricted records and non-archival items. For instance, items related to personnel matters or family medical leave could be redacted, restricted, or purged as appropriate. At the same time, categorized messages would be of immense value to researchers who would no longer have to be as concerned that relevant messages were overlooked in a manual or keyword search.

3. WORKFLOW

3.1 Capstone

The University Archivists have identified approximately 0.1% of its employees as senior administrators for whom most or all email should be retained for its institution-wide value. Another 1% have been identified as mid-level administrators that will frequently have correspondence of significant value to their area of responsibility but do not necessitate retention in bulk.

It is critical to the Capstone approach to inform relevant email account owners of the approach and of the historical value of their correspondence. This opportunity should also be used to address any concerns about the appraisal, transfer, or access restriction processes as well as establish a recurring schedule for ingests. Owners will benefit from specific guidance about items of archival value as well as general email management best practices.

3.2 Transfer

Email transfers frequently occur upon retirement or separation of the individual, which is often when records are most at risk of loss. At a minimum, the office and the successor should retain a copy of important records and correspondence for business continuity purposes.

After a clearly defined period, perhaps 3-6 years after separation, the email should be transferred to the custody of the archives. In a Microsoft Exchange environment, this may be accomplished in a native PST format, possibly using an external hard drive. If possible, custodians should include information describing the main categories of subjects that exist within the correspondence as well as any forms of confidential information that may exist. Custodians may choose to pre-screen the content in order to withhold active or sensitive topics until a later date.

3.3 Processing

3.3.1 Identify

Topics of interest should be identified from transferred email collections. This may be developed through traditional record series and folder lists, sampling informed by the originating office, or using techniques such as data-less classification [10] to gain insights into unknown datasets. De-duplication of identical or nearly identical messages (e.g., sender vs. recipient copy) is also useful at this stage.

3.3.2 Describe

Using a predictive coding tool such as Microsoft's Advanced eDiscovery for Office 365 (formerly Equivio), the messages will be associated with the topics identified above through an iterative training process. Although results may be available through a quick review of as few as 1,000 messages, a greater set of training data will produce more reliable results. Feedback provided during the training process will help determine when training is complete. It is important to note that text must be extracted from the attachments to successfully categorize the document.

3.3.3 Redact

A prioritized review may now be conducted to focus efforts on likely candidates for confidential information. For instance, attorney-client communications and student advising records should be reviewed more carefully while press releases and mass mailings likely require less stringent review. Tools such as Identity Finder or Bulk Extractor may help locate regular forms of personally identifiable information. In addition, review-on-demand services could be offered to provide quick access to researchers while ensuring that access to confidential information is restricted.

3.3.4 Preserve

Multiple tools exist to preserve email, an especially important function given the proprietary and sometimes volatile nature of PST files. Preservica, for instance, uses Emailchemy to extract messages and attachments from PST files and convert the messages to the plain-text EML format. Preservica also supports multiple manifestations, allowing redacted versions of documents in popular formats for public access and un-redacted versions in native and sustainable formats for preservation.

3.4 Access

Although Preservica could also be used to provide online access through its Universal Access feature, many institutions may prefer to maintain offline access using a terminal in the archives. A hybrid of this might utilize the redacted view feature of ePADD [11] to provide limited online keyword search capabilities and general trend visualizations without exposing the full content of a message. Full access may be facilitated in a native email client at the archives terminal. A confidentiality agreement could also be used to further protect against the disclosure of overlooked restricted content.

4. NEXT STEPS

The long-term preservation of digital content presents many challenges to the archival community. The continued custodial responsibilities needed to ensure that content is preserved over time and remains reliably accessible will require thoughtful decisions to be made regarding what content to prioritize. If successful, the use of predictive coding to process Capstone email may provide administrators, researchers, and archivists alike with tools that can assist in making more informed decisions using active and inactive content, responding more swiftly and accurately to requests under freedom of information laws, and performing a limited self-appraisal to identify messages that are of a personal nature or that warrant access restrictions.

During the summer and fall of 2016, the University of Illinois is collaborating with the Illinois State Archives to manually categorize a subset of topics for a 2 million message collection from former Illinois gubernatorial administrations. The results of this effort will be used as part of a National Historical Publications & Records Commission-funded project to evaluate the effectiveness of various predictive coding tools to supplement traditional digital archival methods and ultimately to accession, describe, preserve, and provide access to state government electronic records of enduring value.

5. ACKNOWLEDGMENTS

The authors would like to thank Dave Joens, Kris Stenson, Maura Grossman, Gordon Cormack, Dan Roth, Jordan Phoenix, Shashank Gupta, Roger Christman, NexLP, and the University of Illinois Office of University Relations for their support and collaboration to provide a sustainable path for the digital preservation of email correspondence of enduring value.

6. REFERENCES

[1] University of Illinois Records and Information Management Services. 2015. Preserving Email Messages of Enduring Value. http://go.uillinois.edu/capstone.

[2] U.S. National Archives and Records Administration. 2013. NARA Bulletin 2013-02. http://www.archivess.gov/records-mgmt/bulletins/2013/2013-02.html.

[3] Digital Curation Centre. 2008. DCC Curation Lifecycle Model. http://www.dcc.ac.uk/resources/curation-lifecycle-model.

[4] Greene, M. A. and Meissner, D. 2005. More Product, Less Process: Revamping Traditional Archival Processing. In *The American Archivist,* 68 (Fall/Winter 2005), 208–263. http://www.archivists.org/prof-education/pre-readings/IMPLP/AA68.2.MeissnerGreene.pdf.

[5] University of Illinois Records and Information Management Services. 2014. Metrics Based Reappraisal. http://go.uillinois.edu/rimsMBR.

[6] EDRM. 2014. EDRM Stages. http://www.edrm.net/resources/edrm-stages-explained.

[7] TechTarget. 2010. Electronic discovery (e-discovery or ediscovery). http://searchfinancialsecurity.techtarget.com/definition/electronic-discovery.

[8] RAND Institute for Civil Justice. 2012. Where the Money Goes. http://www.rand.org/content/dam/rand/pubs/monographs/2012/RAND_MG1208.pdf.

[9] Grossman, M. R. and Cormack, G. V. 2016. Continuous Active Learning for TAR. In *Practical Law* (April/May 2016), 32-37. http://cormack.uwaterloo.ca/cormack/caldemo/AprMay16_EdiscoveryBulletin.pdf.

[10] University of Illinois Cognitive Computation Group. 2014. Dataless Classification. https://cogcomp.cs.illinois.edu/page/project_view/6.

[11] Stanford University. 2015. ePADD. https://library.stanford.edu/projects/epadd.

Using RMap to Describe Distributed Works as Linked Data Graphs: Outcomes and Preservation Implications

Karen L. Hanson
Sheila Morrissey
Portico
100 Campus Drive, Suite 100
Princeton, NJ 08540
+1 609-986-2282
karen.hanson@ithaka.org
sheila.morrissey@ithaka.org

Aaron Birkland
Tim DiLauro
Johns Hopkins University
3400 N. Charles Street / MSEL
Baltimore, MD 21218
+1 410-929-3722
apb@jhu.edu
timmo@jhu.edu

Mark Donoghue
IEEE
445 Hoes Lane
Piscataway, NJ 08854
+1 732-562-6045
m.donoghue@ieee.org

ABSTRACT

Today's scholarly works can be dynamic, distributed, and complex. They can consist of multiple related components (article, dataset, software, multimedia, webpage, etc.) that are made available asynchronously, assigned a range of identifiers, and stored in different repositories with uneven preservation policies. A lot of progress has been made to simplify the process of sharing the components of these new forms of scholarly output and to improve the methods of preserving diverse formats. As the complexity of a scholarly works grows, however, it becomes unlikely that all of the components will become available at the same time, be accessible through a single repository, or even stay in the same state as they were at the time of publication. In turn, it also becomes more challenging to maintain a comprehensive and current perspective on what the complete work consists of and where all of the components can be found. It is this challenge that makes it valuable to also capture and preserve the map of relationships amongst these distributed resources. The goal of the RMap project was to build a prototype service that can capture and preserve the maps of relationships found amongst these distributed works. The outcomes of the RMap project and its possible applications for preservation are described.

Keywords

Publishing workflows; linked data; data publishing; semantic web; RESTful API; digital preservation; scholarly communication; digital scholarship.

1. BACKGROUND

In recent years, the content that comprises the scholarly record has shifted from being primarily discrete text-based bounded objects, such as journals or books, to more dynamic and less "bounded" content that might include data, webpages, software, and more. In other words, the boundaries of the scholarly record are stretching beyond the traditional publication of outcomes to instead encompass additional outputs created during the process and aftermath of the work [10]. This means a scholarly work can be complex, dynamic, and consist of multiple distributed parts. An example of a typical map of the heterogeneous resources that comprise and describe a single work is shown in Figure 1.

These changes in scholarly communication have been facilitated by technological shifts that have diversified the kinds of content that can be produced during research and made it easier to share digital material. One consequence of this has been a movement towards more funders and publishers requesting that researchers maintain and/or share research outputs to support reuse, validation, and replication of their methods and results. In the US, for example, the Office of Science and Technology Policy's 2013 memorandum [8] highlighted the government's commitment to improving availability of data resulting from federally funded research. An example in publishing is Nature Publishing Group's policies requiring that authors make materials, data, code, and protocols available to readers on request[1].

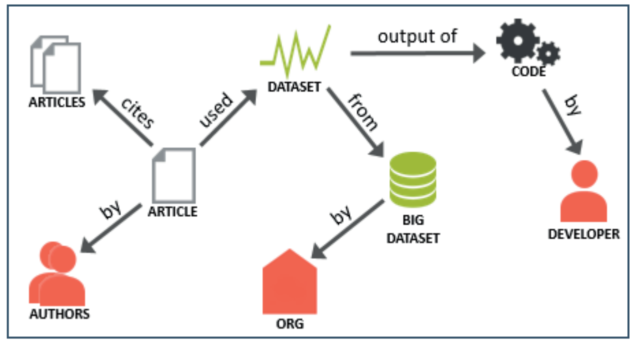


Figure 1 Multi-part Distributed Scholarly Work

Another consequence is the changes to publication workflows to support citing these new kinds of materials. For example, a lot of work has been done to support citing datasets in articles as first-class objects. Other initiatives have expanded this effort to include software citation [1] and citation of other kinds of resources, such as antibodies or model organisms [2]. Guidelines on data citation have been implemented by some publishers, though they are not yet consistently applied. One study shows only 6% of Dryad datasets associated with a journal article appear in the citation list for that article [11].

While this expansion of categories of citation is useful, there are many shortcomings attendant on attempting to shoehorn the rich network model of scholarly artifacts, contexts, and relationships into the structure of a journal article citation. First is the challenge inherent in the asynchronous nature of publishing the various components of a distributed work. Once an article is published in the traditional manner, the opportunity to connect or update related works has often passed, or at least become more difficult, with incentives for the author to update the connections greatly reduced. In an ideal scenario, all supporting material and outputs would be published and assigned identifiers before an article is published, but in reality this can be difficult to orchestrate and happens rarely. This means additional context shared post publication cannot typically be referenced from the published article. Second, the OCLC report on *The Evolving Scholarly Record* [10] describes how even after research outcomes are published, useful context and commentary are added to the work through presentations, blogs, and more in the "aftermath." These responses may never be published in an article with a DOI, but can provide important context to the work. A third challenge is that, while some publishers accept many kinds of works in their citation list, others are more restrictive. There are some works that cannot be published in a repository or easily assigned an identifier because their dynamic nature, scale, or copyright. If citations lists are limited to items with certain kinds of identifiers, for example, some components may not be included. Fourth, publisher or editorial boards often limit the total number, not just the type, of citations. Furthermore, there are sometimes simply too many objects for traditional citation to be practical. Finally, a linear citation list may not allow the researcher to clearly capture the role a resource played in the work or the nature of various contributions to the project.

All of these challenges suggest that there could be value in a service that can capture and preserve these evolving maps of relationships among the resources that form the scholarly work. One of the important tenants of the RMap Project is that this map itself can be considered a first class artifact of scholarly communication. For an increasing number of works, the published article is the tip of the iceberg. The definition of what encompasses a scholarly work has become much more complex than it once was. Understanding how the parts of a work relate to each other is important context for being able to preserve scholarship in a way that will allow it to be reused, replicated and validated.

2. THE RMAP PROJECT

The RMap[2] project was funded by the Alfred P. Sloan Foundation[3] and carried out by the Data Conservancy[4], Portico[5], and IEEE[6], starting in 2014. The goal of the project was to create a prototype API that could capture and preserve the maps of relationships amongst scholarly works.

The RMap team's work was developed in the context of a growing consensus that there is a need to capture the relationships amongst the components of complex scholarly works. The OCLC report on *The Evolving Scholarly Record* [10] identified the need for the expression of a set of relationships to bind together the pieces of a scholarly work. The Research Object collaboration has produced a set of tools and specifications for bundling together and describing essential information relating to experiments and investigations [3]. The RDA/WDS Publishing Data Services Working Group, in which the RMap team has participated, recently published recommendations for implementing a data to publication cross-linking service [5]. The working group also implemented a pilot aggregation and query service[7] and continue to develop the framework under the name *Scholix*[8]. More recently DataCite announced its Event Data service[9], which will support the registration and exchange of references between resources.

Some of these services focus on bilateral connections between objects, often with a circumscribed set of defined relationships between objects, and with allowable persistent identifiers for resources. RMap's focus is on the complete graph of resources that represent a compound work, with support for all identifiers and relationships that can be expressed as valid linked data. Through these graphs, bilateral relationships can also be identified. Over the last 2 years the RMap project has developed an API service that can act as a hub for capturing and preserving these maps.

RMap captures the resource maps as linked data[10] graphs, building on the features of the semantic web [4] and adopting the concept of an Aggregation from the Open Archives Initiative Object Reuse and Exchange[11] (OAI-ORE) standard. To support easy integration into existing data workflows, RMap employs a RESTful (Representational State Transfer) API [6]. Where available, RMap makes use of existing broadly adopted vocabularies (e.g. Dublin Core[12], Friend of a Friend[13], Open Provenance Model[14]) in its data model.

2.1 Objectives

As we have noted, RMap aims to capture and preserve links amongst the artifacts of scholarly communication and those who create, modify, employ, and annotate them [7]. Its purpose in doing so is to facilitate the discovery and reuse of those artifacts, to demonstrate the impact and reuse of research, to make those demonstrations available to those making curatorial decisions about collection and preservation of digital research artifacts such as software and workflows, and to inform those curatorial and other choices with solid provenance information about the assertions recorded in RMap.

Key design objectives of the RMap service in support of these goals are to

- support assertions from a broad set of contributors
- integrate with Linked Data
- leverage existing data from other scholarly publishing stakeholders (publishers, identifier providers, identity authorities, data, and software repositories)
- provide some support for resources lacking identifiers

2.2 Data Model

The RMap data model utilizes the Resource Description Framework (RDF)[15] concepts of resources, triples, and graphs. The model includes three kinds of named graphs: *DiSCOs*, *Agents*, and *Events*.

2.2.1 RMap DiSCOs

RMap DiSCOs (Distributed Scholarly Compound Objects) are named graphs containing:

- A unique persistent identifier
- A list of 1 or more aggregated resource URIs (*ore:aggregates*) that form the aggregated work.
- An optional list of assertions about the aggregated resources. There are no constraints on the ontologies that can be used in these assertions, provided they form a connected graph with the aggregated resources at the root. These may be used to include additional context about each of the resources e.g. descriptive metadata, relationships to other resources, type, other identifiers, etc.
- An optional creator, description, and provenance URI to provide more information about the source of the DiSCO.

[1] http://www.nature.com/authors/policies/availability.html

[2] http://rmap-project.info

[3] http://www.sloan.org/

[4] http://dataconservancy.org/

[5] http://www.portico.org/

[6] http://www.ieee.org/

[7] http://dliservice.research-infrastructures.eu/

[8] http://www.scholix.org/

[9] https://eventdata.datacite.org/

[10] http://www.w3.org/standards/semanticweb/data

[11] http://www.openarchives.org/ore/

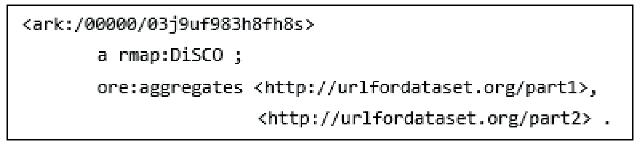
[12] http://dublincore.org/specifications/

[13] http://xmlns.com/foaf/spec/

[14] http://openprovenance.org/

[15] http://www.w3.org/TR/rdf11-concepts/

DiSCOs contain the ore:aggregates predicate, but do not otherwise follow the OAI-ORE model. For example, while OAI-ORE logically separates the concept of an Aggregation from the document that describes it (the "Resource Map"), a DiSCO combines these two notions into a single resource in order to make it easier to contribute data. Instead, much of the data that would typically be part of an OAI-ORE Resource Map is generated automatically as part of the API service and stored as RMap Events. As a result, the simplest form of a DiSCO is very easy to construct. An example of this is shown in Figure 2, which simply asserts that two resources form a compound object but does not further define the relationship between them. Beyond this users can add as much detail to the DiSCO as they see fit. The RMap team chose to keep the model simple and requirements to a minimum, but have also investigated what would be required to make the system fully compatible with OAI-ORE. It is estimated that the OAI-ORE model could be supported with several small enhancements if there were demand for this in the future.



```
<ark:/00000/03j9uf983h8fh8s>
        a rmap:DiSCO ;
        ore:aggregates <http://urlfordataset.org/part1>,
                       <http://urlfordataset.org/part2> .
```

Figure 2 Simple DiSCO as Turtle RDF

DiSCOs are immutable in that their identifier always corresponds to a specific set of assertions. When a DiSCO is updated, the previous version still exists and the new version is assigned a new identifier.

DiSCOs can have one of four statuses. *Active* means the assertions in the DiSCO are still assumed to be true. *Inactive* means the DiSCO has either been retracted or updated with a new set of assertions. Inactive DiSCOs can still be accessed publicly. When a DiSCO is updated, the previous version is automatically set to Inactive, but is still available to view in the version chain. *Deleted* means the DiSCO is retracted and the assertions are not publicly visible through the API, even though the data exists in the database. A *Tombstoned* status means the DiSCO has been removed from the database, but the provenance information persists as a record of the removal.

2.2.2 RMap Agents

RMap *Agents* are named graphs representing a person, process, or thing that is responsible for some action on the RMap database. Anyone who contributes data to RMap is required to have an Agent. Each new Agent is assigned a persistent identifier that is associated with changes to the database. Unlike the DiSCO model, Agents are mutable, so updates to the Agent graph will overwrite the previous version. Changes to the Agent graph are recorded as *Events*.

2.2.3 RMap Events

An RMap *Event* is automatically generated whenever a user makes any additions or changes to RMap. They are used to record and track the provenance and status of RMap DiSCOs and Agents. Each Event has a unique persistent identifier, and includes the URI of the RMap Agent that made the change, the type of change, URIs of any RMap objects affected, the timeframe of the Event, and optionally the specific API access key that was used to make the change. Events cannot be updated or deleted.

2.3 RESTful API

The primary interface for accessing the RMap database is a RESTful API. The features of a RESTful API include programming language independence and conformance to web architecture metaphors. Both are important in facilitating the integration of the RMap service into heterogeneous publisher, researcher, funder, and other institutional workflows.

The RMap RESTful API includes over 30 functions for querying and generating data. For example, you can retrieve a list of triples that mention a specific resource, or a list of DiSCOs created by a specific Agent. Functions that generate lists of results can typically be filtered by date, creating Agent, and DiSCO status.

2.4 Web Application and Visualization Tool

In addition to the RESTful API, data can be navigated interactively through the RMap web application. This allows the user to look up DiSCO URIs and view either a tabular representation or a graph visualization (Figure 3) of the data. By clicking on resources in the visualization or data table, it is possible to drill into the data and view all triples and DiSCOs that reference that resource.

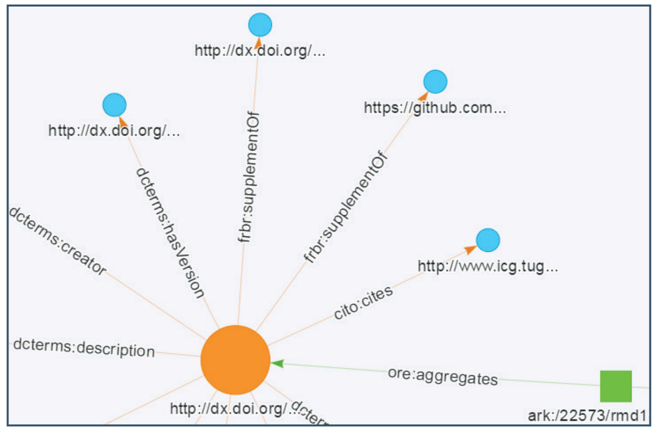


Figure 3 Part of RMap DiSCO visualization

2.5 Outcomes

Over the last two years, the RMap team has produced a working prototype RESTful API for managing and retrieving data in RMap. They have also built a web application for navigating the RMap data interactively. By logging into the web application using Twitter, ORCID, or Google authentication, users can generate keys for the RESTful API. Links to the tools and documentation can be found on the RMap website[16]. Also available is a versatile harvesting framework to support large scale harvesting and ingest of DiSCOs. The team has also explored options for an inferencing engine to support the mapping of equivalent identifiers.

Example DiSCOs were created using metadata from DataCite[17], NCBI's PubMed and Nuccore APIs[18], ACM's Transactions of Mathematical Software[19], Portico, and the complete collection of IEEE articles. In one example metadata relating to a single article was imported from IEEE, Portico, and DataCite in order to demonstrate how to navigate between different components of a work through overlapping DiSCOs. The RMap database continues to grow. At the time of writing the RMap prototype service contains over 4.5 million DiSCOs, comprised of over 230 million triples.

[16] http://rmap-project.info

[17] https://www.datacite.org/

[18] http://www.ncbi.nlm.nih.gov/home/api.shtml

[19] http://toms.acm.org/

A short extension to the project is supporting the exploration of representing SHARE[20] and Open Science Framework[21] data as DiSCOs in RMap.

3. PRESERVATION IMPLICATIONS

The goal of the RMap project was to develop a framework for capturing and *preserving* maps of relationships. Since RMap DiSCOs can be exported as RDF text format, exporting and preserving RMap DiSCOs can follow a typical preservation pathway for plain text. As the project has unfolded, however, some other potential preservation use cases have been identified.

While the pathways to preservation of articles produced by publishers are well understood, the other components of the scholarly works described previously are typically not preserved in the same repository. Even if all of the components of the work are available in other repositories, it is unlikely that the map of the connections between all of the parts will be available in a form that is accessible to all repositories. This means none of the components show a full picture and the complete work is difficult to assemble. Using RMap as a hub to represent these connections between the distributed components of the works, could help ensure all components of the work can be found and preserved.

Where metadata and components are distributed across different kinds of platforms, it is possible that one or more of the resources will eventually be lost or altered. Even if all resources are preserved, it is highly likely that one of the resources will reference a URL that has moved or no longer exists and will produce a 404 "not found" error when accessed. One study showed that the problem of *reference rot* already affects one in five articles [9]. Add to that equation a variety of non-article resources that are not necessarily peer reviewed or conforming to any fixed publication path, and the problem of reference rot may be even more problematic. Even if there is a new equivalent link available, there is often no easy way for anyone to indicate a new location. Not only does RMap provide an opportunity for links to be updated and identifiers added, one useful enhancement to the framework might be to interface with the Internet Archive's Wayback Machine[22] APIs to associate Memento links with web URLs that do not use a persistent URI.

Finally, during the first phase of the project, the RMap team generated some DiSCOs using Portico content. Each DiSCO showed which resources were preserved by Portico for a single article. Combining similar data from other repositories could be useful for identifying preservation gaps and overlap for different kinds of work.

4. CONCLUSIONS

The RMap project has produced a framework for generating maps of the components of a distributed scholarly work. By being part of publisher, researcher, funder, and other scholarly workflows and by aggregating data from multiple sources, RMap aims to support third party discovery as well as facilitate the capture of information about scholarly artifacts that is not easily captured elsewhere. Some applications of RMap could also support improved preservation of distributed scholarly compound works.

5. ACKNOWLEDGMENTS

The RMap Project is funded by the Alfred P. Sloan Foundation. The authors wish to acknowledge the contributions of their RMap project colleagues: Sayeed Choudhury, Johns Hopkins University, The Sheridan Libraries, Associate Dean for Research Data Management; Kate Wittenberg, Managing Director, Portico; Gerry Grenier, Senior Director, Publishing Technologies, IEEE; Portico colleagues Jabin White, Vinay Cheruku, Amy Kirchhoff, John Meyer, Stephanie Orphan, Joseph Rogowski; and IEEE colleagues Renny Guida, Ken Rawson, Ken Moore.

6. REFERENCES

[1] Ahalt, S., Carsey, T., Couch, A. et al. 2015. NSF Workshop on Supporting Scientific Discovery through Norms and Practices for Software and Data Citation and Attribution. Retrieved April 2016 from https://softwaredatacitation.org/Workshop%20Report

[2] Bandrowski, A., Brush, M., Grethe, J. S. et al. 2015. The Resource Identification Initiative: A cultural shift in publishing [version 2; referees: 2 approved]. *F1000Research* 4, 134. DOI= http://doi.org/10.12688/f1000research.6555.2.

[3] Bechhofer, S., Ainsworth J., Bhagat, J. et al. 2013. Why Linked Data is Not Enough for Scientists. *Future Generation Computer Systems* 29, 2 (February 2013), 599-611. DOI= http://doi.org/10.1016/j.future.2011.08.004

[4] Berners-Lee, T., Hendler, J. and Lassila, O. 2001. The semantic web. *Scientific American*, *284*(5), 28-37.

[5] Burton A., and Koers, H. 2016. *Interoperability Framework Recommendations.* ICSU-WDS & RDA. Publishing Data Services Working Group. Retrieved 29 June 2016 from http://www.scholix.org/guidelines

[6] Fielding, R. T. 2000. *Architectural Styles and the Design of Network-based Software Architectures*. Dissertation. University of California, Irvine. Retrieved 26 January 2015 from https://www.ics.uci.edu/~fielding/pubs/dissertation/fielding_dissertation.pdf

[7] Hanson, K. L., DiLauro, T. and Donoghue, M., 2015. The RMap Project: Capturing and Preserving Associations amongst Multi-Part Distributed Publications. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '15). ACM, New York, NY, USA, 281-282. DOI= http://dx.doi.org/10.1145/2756406.2756952

[8] Holdren, J.P., 2013. *Increasing access to the results of federally funded scientific research. Memorandum for the heads of executive departments and agencies.* Office of Science and Technology Policy, Executive Office of the President, Washington, DC. Retrieved 20 April 2016 from https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[9] Klein, M., Van de Sompel, H., Sanderson, R., et al. 2014. Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot. *PLoS ONE,* 9, 12. e115253. DOI= http://doi.org/10.1371/journal.pone.0115253

[10] Lavoie, B., Childress, E., Erway, R., Faniel, I., Malpas, C., Schaffner, J. and van der Werf, Titia. 2014. *The Evolving Scholarly Record*. OCLC Research, Dublin, Ohio. Retrieved 20 April 2016 from http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-evolving-scholarly-record-2014.pdf

[11] Mayo, C., Hull, E.A. and Vision, T.J. 2015. The location of the citation: changing practices in how publications cite original data in the Dryad Digital Repository. *Zenodo.* DOI= http://dx.doi.org/10.5281/zenodo.32412

[20] http://www.share-research.org/

[21] https://osf.io/

[22] http://archive.org/web

LONG PAPERS //

An OAIS-oriented System for Fast Package Creation, Search, and Access

Sven Schlarb, Rainer Schmidt, Roman Karl, Mihai Bartha, Jan Rörden
AIT Austrian Institute of Technology
Donau-City-Straße 1
1220 Vienna, Austria
{first}.{last}@ait.ac.at

Janet Delve
University of Brighton
CRD, Grand Parade
Brighton, BN2 0JY, UK
BN2 4AT
J.Delve@brighton.ac.uk

Kuldar Aas
National Archives of Estonia
J. Liivi 4
Tartu, 50409, Estonia
kuldar.aas@ra.ee

ABSTRACT

This paper describes the scalable e-archiving repository system developed in the context of the E-ARK project. The system is built using a stack of widely used technologies that are known from areas such as search engine development, information retrieval and data-intensive computing, enabling efficient storage and processing of large volumes of data. The E-ARK Integrated Platform Reference Implementation Prototype takes advantage of these technologies and implements an OAIS-oriented repository system for creating, archiving, and accessing data as information packages. The system consists of software components including an efficient file handling infrastructure, a configurable and scalable ingest system, a powerful full-text-based search server, and a distributed repository providing file-level random access. This paper gives an overview of the architecture and technical components that have been used to build the prototype. Furthermore, the paper provides experimental results and gives directions for future work.

Keywords

OAIS; archiving; repository; scalability; distributed systems; Hadoop

1. INTRODUCTION

In recent years, considerable research and development efforts dealt with managing the growing amount of digital data that is being produced in science, information technology, and many other areas of today's society [9]. The constant increase in the number of digital publications, governmental records, or digitized materials is challenging for the development of procedures and information systems for libraries and archives [10]. An effort to cope with preservation workflows that need to be executed on large volumes of digital materials has been made by the SCAPE project [12], which has developed a platform that enables users to execute such processes using computer clusters and data-intensive computing techniques [19]. The E-ARK Integrated Platform Reference Implementation Prototype[1] continues this work by setting up a scalable repository system for archival institutions.

The integrated prototype has been developed in the context of the E-ARK project, an ongoing 3-year multinational research project co-funded by the European Commission's ICT Policy Support Program (PSP) within the Competitiveness and Innovation Framework Program (CIP). The purpose of the integrated prototype is to demonstrate how open source solutions for distributed storage and processing can be combined to build a scalable repository for archiving organizations. The aim is to show that this approach is, in general, suitable to address the need for enhancing existing archiving systems in providing access to very large, continuously growing, and heterogeneous digital object collections in archival institutions.

In its first project year, E-ARK has conducted a GAP analysis among archival institutions identifying user requirements for access services[2]. The study investigated the current landscape of archival solutions regarding the available access components and identified gaps and requirements from the perspective of national archives, 3rd party users, as well as content providers. The study identified a major gap in the identification process where users browse and search collections to identify material of potential interest. It stated that a lack of comprehensive metadata available and indexed compromises the performance and efficiency of the finding aids, which directly impacts the user experience and the user's access to the archival holdings in their entirety.

To fill this gap, E-ARK makes use of s scalable repository system and search infrastructure for archived content. The goal is not necessarily to replace existing systems but to augment these components (like archival catalogues) with a "content repository" that can be searched based on a full text index. The content repository concentrates on fine grained search within information packages and random access at the file-level rather than providing search based on selected metadata elements and package-based access. The integrated prototype developed in this context employs scalable (cluster) technology as scalability issues must be taken into account when operating a detailed content-based search facility, providing an infrastructure for creating, ingesting, searching, and accessing E-ARK information packages. Scalability is accomplished by making use of technologies like the

[1] in the following shortly called "integrated prototype".

[2] http://www.eark-project.com/resources/project-deliverables/3-d51-e-ark-gap-report

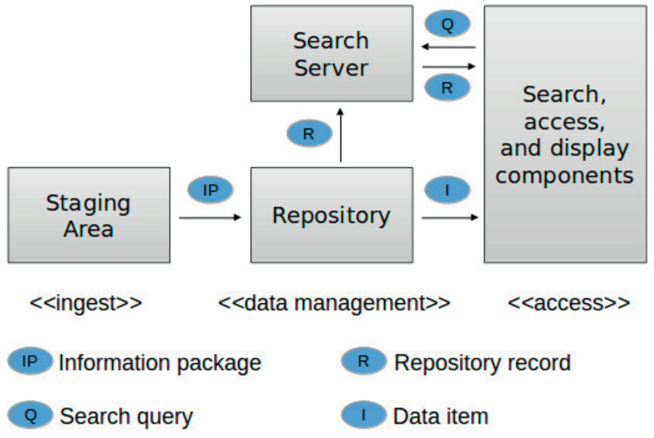


Figure 1: System Components and their interactions used by the integrated prototype for implementing the Faceted Query Interface and API.

Apache Hadoop framework[3], NGDATA's Lily repository[4], and the Apache SolR search server[5].

The workflow implemented by the integrated prototype for data ingest, storage, and access is based on the ISO Reference Model for an Open Archival Information System (OAIS) [3]. This means that data is received in form of Submission Information Packages (SIPs) which are transformed into Archival Information Packages (AIPs) and transferred to the archive. Upon client request the selected content of the AIPs can be retrieved from the archive and repacked as Dissemination Information Packages (DIPs) for delivery. The repository supports facet search based on full-text and extracted metadata (e.g. MIME-type, size, name of the files contained within the information packages). This is accomplished by executing information extraction and transformation processes upon transfer of the SIP to the archive (SIP to AIP conversion/ingest).

2. BACKEND ARCHITECTURE

2.1 Overview

Figure 1 provides an overview of the major system components that are employed by the backend of the integrated prototype. The query interface and API provided by the search server must be backed by software components and generated data products in order to provide the desired functionality. Here, we give an overview and describe their interactions.

2.2 Staging Area

The staging area is a file-system based storage location provided in combination with the data management component of the integrated prototype. The staging area is accessible to other components based on an API allowing these components to deposit information packages for ingestion into the content repository (as shown in Figure 1). While in principle any file system could be employed as staging area, the integrated prototype makes use of the Hadoop File System (HDFS) for performance, scalability and reliability reasons. The staging area is in the first place used to access the information packages during the repository ingest workflow but can also be employed to serve other purposes like archival storage and (package-based) access.

2.3 Repository

The integrated prototype makes use of NGDATA's Lily project which is employed as a content repository. The information packages residing on the staging area are ingested into the repository where they are stored in the form of structured repository records, as described in section 3. The repository interacts with the search server which reads and indexes the repository records as well as with client components which access data items on a file or record level.

2.4 Search Server

The generation and/or update of the index provided by the search server can be triggered by the repository component in case records are added, deleted, or modified. The index provides the necessary data structure to evaluate search queries and to return results which point to records stored in the repository. The index and search functionality is provided by the search server through an HTTP interface. The integrated prototype makes use of Apache Solr as the search server which can be well integrated with Lily and its underlying database HBase. The query interface is provided by a defined REST API through Apache Solr which is customized based on the individual structure of the repository records. For supporting multiple and heterogeneous collections, it is possible to generate different indexes for different datasets maintained by the repository.

2.5 Search, Access, and Display Components

These components interact with the search server and the repository as clients. Specific archival user interface and access components (e.g. required for DIP creation) have been implemented in the context of the E-ARK Web project, as described in section 5.2. The protocol for interacting with the query interface is however independent of the employed client component and ultimately allows for the integration with an external user interface. Client components typically provide a graphical representation of the query language and facets provided by the search server. When a query is submitted to the search server, it is evaluated against the index. The search server subsequently returns a ranked list of record references (and optionally content fragments) to the client. Besides interfaces required for searching, the repository also provides an access service providing clients with random access to data on a file-level, based on references, which can retrieved by an HTTP request, issued for example through the client application.

3. CONCEPTUAL WORKFLOW

Figure 2 shows the conceptual workflow for ingesting data items residing on the staging area (for example using the Hadoop File system) into the content repository. Practically, this means that after the repository has been populated and/or updated a full text index is generated and/or updated respectively.

The integrated prototype implements the ingest workflow for ingesting information packages into the content repository on a file-based level which is in contrast to ingesting on a package level. The ingest workflow is implemented in a way that every item (or file) contained within an information package is considered a record. In the repository, packages are represented as a set of records sharing a common identifier.

[3] http://hadoop.apache.org/
[4] https://github.com/NGDATA/lilyproject
[5] http://lucene.apache.org/solr/

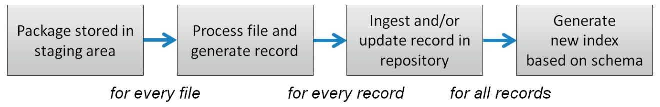

Package stored in staging area
Process file and generate record
Ingest and/or update record in repository
Generate new index based on schema
for every file
for every record
for all records

Figure 2: Conceptual workflow for ingestion and indexing of information packages to the content repository provided by the integrated prototype.

3.1 Record Extraction and Ingest

Once the ingest process is started, the workflow iterates over all files contained within the individual information packages. Each file extracted from the information package is processed separately. The exact implementation of the processing step is highly depending on the data set and institutional requirements. Examples that have been implemented as part of the integrated prototype include the extraction of text portions, structure, and context information from web, office, or XML documents, file size calculation, MIME-type identification, and checksums.

The data extracted from an individual file is subsequently stored into a data structure, called a record , which can be ingested into the repository. The individual structure of a record can be customized depending on data and institutional needs. A record for a newspaper article, for example, could contain fields like author, title, body, publisher, and publishing date. Fields of a record are "typed" which means they can be restricted to certain data types like for example numbers, string, or date. A record identifier that encodes the identifier of the package as well as the location of the original file within the package is generated automatically. Once a record is created, it is ingested into the content repository. As records organize information in a structured way, they can be interpreted by the repository and consequently stored in a (structured) database.

3.2 Full-Text Index Generation

The E-ARK integrated prototype aims at providing a facet query interface based on full-text indexing in addition to the rather limited search mechanisms provided through database indexing. The full-text search functionality is provided through a search server (like Apache Solr), which relies on a previously generated full-text index (using Apache Lucene). The integrated prototype makes use of a configuration file (called a schema) that provides a detailed specification of the indexing process. This controls for example which parts of a record should be indexed, available fields, and the information that should be stored with the index (e.g. only document references and/or also content portions).

After new content has been ingested and/or updated the repository index should be generated or updated at periodic intervals. The integrated prototype provides specific commands for triggering the creation of the index from the records available within the repository. Depending on the volume of content, indexing as well as ingestion can become very resource and time consuming processes. Both processes have therefore been implemented as parallel applications that can take advantage of a computer cluster to scale out for large data sets. Within the E-ARK integrated prototype, indexing and ingestion have been deployed on a cluster at AIT, providing a total of 48 CPU-cores. The generated index is made available by the search server as a query interface enabling a client to formulate and execute queries against the index, compose complex queries based on facets, and rank them based on different characteristics. It is however important to note that although a defined query API is exposed by the integrated prototype, the API is highly configurable and customizable with respect to the parameters it accepts and the nature of results it returns.

Table 1: Daemons running on the cluster.

	Master	*Slave*
HDFS	NameNode	DataNode
MapReduce	JobTracker	TaskTracker
HBase	HBase Master	Region Server

The workflow shown in Figure 2 was implemented based on the software components described in section 2 (and shown in Figure 1). It has been configured for different test data and deployed in a single-node environment as well as in a cluster environment available at AIT.

4. SCALABLE PROCESSING AND SEARCH INFRASTRUCTURE

Although systems for parallel and distributed computing have been studied since the early 1980's and parallel database systems were established already in the mid-1990's [1], a significant change in the last decade occurred with the advent of the MapReduce data processing paradigm [5] and the subsequent rise of open source technology for distributed storage and parallel data processing provided by Apache Hadoop. In the following, we describe the integrated prototype backend which is based on Apache Hadoop and related components that emerged in the Hadoop ecosystem during the last decade.

4.0.1 Hadoop

The backend system of the integrated prototype is built on top of the Hadoop framework and can be deployed on a computer cluster allowing the repository infrastructure to scale-out horizontally. This enables system administrators to increase the available system resources (i.e. for storage and processing) by adding new computer nodes. Using Hadoop, the number of nodes in a cluster is virtually unlimited and clusters may range from single node installations to clusters comprising thousands of computers.

Usually one would, however, build a cluster consisting of a master node and at least two slave nodess to get a performance advantage from the distributed environment. Each slave machine runs all services, which means that it runs a DataNode, a TaskTracker and a Region Server. For production clusters, it is recommended to deploy the NameNode on its own physical machine and furthermore use a SecondaryNameNode as a backup service. Although Lily is deployed on multiple nodes, it does follow the concept of master and slave nodes. There is only one type of Lily node which is intended to run co-located with Region Servers on the cluster.

4.0.2 *Lily*

Lily provides a repository that is build on top of HBase, a NoSQL database that is running on top of Hadoop. Lily defines some data types where most of them are based on existing Java data types. Lily records are defined using these data types as compared to using plain HBase tables, which makes them better suited for indexing due to a richer data model. The Lily Indexer is the component which sends the data to the Solr server and keeps the index synchronized with the Lily repository. Solr neither reads data from HDFS nor writes data to HDFS. The index is stored on the local file system and optionally distributed over multiple cluster nodes if index sharding or replication is used. Solr can be run as a standalone Web-based search server which uses the Apache Lucene search library for full-text indexing and search. The integrated prototype utilizes the Lily Java API as part of a Hadoop MapReduce job in order to ingest large volumes of files in parallel.

4.0.3 *Solr*

There are several options to run Solr. The first option is to run Solr only on one machine. In this case the index is not split and only one shard is used. The second option is to use multiple shards and configure Lily to distribute the input over all shards. As Solr 4 introduced SolrCloud, this became the third option, and it is also the preferred option for a production system. SolrCloud does not only take care of the sharding, it also provides a mechanism for replication. Using Lily in combination with SolrCloud requires some additional configuration work being done, as Lily was developed against Solr 4.0, where SolrCloud was not yet entirely mature. For an example, it is required to create an empty directory in ZooKeeper manually where SolrCloud can store its information.

4.0.4 *ZooKeeper*

HBase, but also Lily and SolrCloud, depend on a running ZooKeeper cluster. ZooKeeper is a framework that supports distributed applications in maintaining configuration information, naming, providing distributed synchronization, and providing group services. ZooKeeper stores small amounts of information, typically configuration data, which can be accessed by all nodes. For experimental clusters that do not need to provide high fault tolerance, it is sufficient to run one ZooKeeper node, which is also called *Quorum Peer*. A higher fault tolerance can be achieved by running three, five or more Quorum Peers. If more than half of the nodes keep running without failures, ZooKeeper stays reliable.

5. FRONTEND ARCHITECTURE

5.1 Overview

In general, the backend system of the integrated prototype takes information packages as input and provides functionalities like information extraction, search, and random access for the contained data items. In the previous chapters, we have outlined a set of custom components and services which have been specifically developed to realize the integrated prototype. The described E-ARK Web Project provides a lightweight front-end implementation for this backend system. The responsibility of the frontend system is the provisioning of user interfaces and corresponding services for creating information packages like AIPs and DIPs.

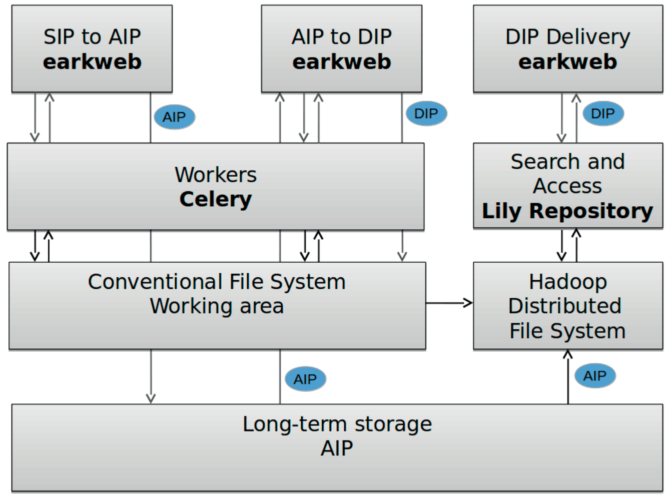


Figure 3: The architecture consists of user interface components that support information package ingest and access processes. The frontend components are backed by a package creation infrastructure handling file, task, and workflow processing. The frontend system is integrated with the Hadoop backend infrastructure for content extraction, storage, and search.

The architecture consists of user interface components that support information package ingest and access processes. The frontend components are backed by a Package Creation Infrastructure handling file, task, and workflow processing. The frontend system is integrated with the Hadoop backend infrastructure for content extraction, storage, and search. The implementations provided by the integrated prototype are light weight applications which are limited in their functionality and focused on distinct E-ARK principles. The architecture of the integrated prototype is in general designed to support a loose coupling strategy so that existing systems can be combined with and/or be augmented with the integrated prototype or particular components of the integrated prototype platform.

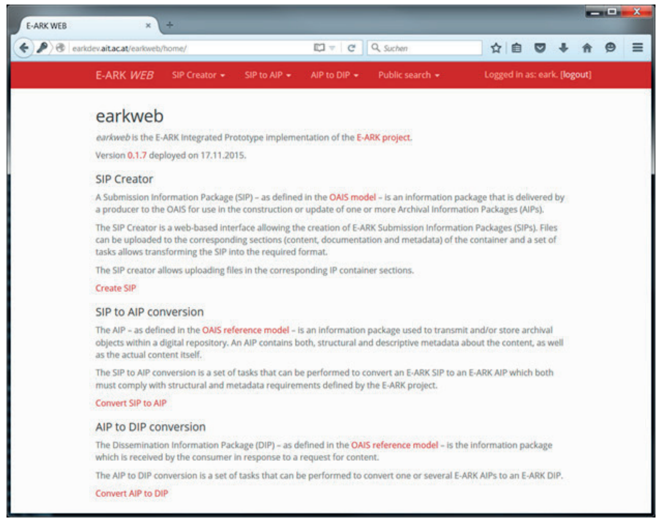
5.2 The E-ARK Web Project

The project *E-ARK Web* [6] is a web application together with a task execution system which allows synchronous and asynchronous processing of information packages by means of processing units which are called "tasks". The purpose of E-ARK Web is, on the one hand, to provide a user interface for the integrated prototype in order to showcase archival information package transformation workflows which are being developed in the E-ARK project in an integrated way. On the other hand, the goal is to provide an architecture which allows reliable, asynchronous, and parallel creation and transformation of E-ARK information packages (E-ARK SIP, AIP, and DIP) integrated with E-ARK backend services for scalable and distributed search and access.

The components of the E-ARK Web project coordinate package transformations between the package formats SIP, AIP, and DIP, and uses Celery [7], a distributed task queue, as its main backend, shown in figure 3. Tasks are designed to perform atomic operations on information packages and any dependency to a database is intentionally avoided to increase processing efficiency. The outcome and status of a task's process is persisted as part of the package. The E-ARK Web project also provides a web interface that allows one to orchestrate and monitor tasks by being loosely coupled with the backend. The backend can also be controlled via remote command execution without using the web frontend. The outcomes of operations performed by a task are stored immediately and the PREMIS format [2] is used to record digital provenance information. It is possible to introduce additional steps, for example, to perform a roll-back operation to get back to a previous processing state in case an error occurs.

5.3 The E-ARK Web User Interface

The user interface of the integrated prototype is a Python[8] /Django[9]-based web application which allows for managing the creation and transformation of E-ARK information packages (E-ARK IPs). It supports the complete archival package transformation pipeline, beginning with the creation of the Submission Information Package (SIP), over the conversion to an Archival Information Package (AIP), to the creation of the Dissemination Information Package (DIP) which is used to disseminate digital objects to the requesting user. The E-ARK Web website is divided into four main areas: First, there is the "SIP creator" area which allows initiating a new SIP creation process and offers a set of transformation tasks to build E-ARK compliant SIPs. Second, there is the "SIP to AIP" area that allows for the execution of tasks for converting an E-ARK compliant SIP to an AIP. Third, there is the "AIP to DIP" area which allows initiating a DIP creation process based on previously selected AIPs used for building the DIP with the help of a set of corresponding conversion tasks. And, finally, there is the "Public search" area offering full-text facet search based on the textual content available in the AIPs which have been uploaded to the HDFS staging area, ingested into Lily, and full-text indexed using SolR, as described in section 3. A screenshot of this user interface is shown in Figure 4.



earkweb

earkweb is the E-ARK Integrated Prototype implementation of the E-ARK project.

Version 0.1.7 deployed on 17.11.2015.

SIP Creator

A Submission Information Package (SIP) – as defined in the OAIS model – is an information package that is delivered by a producer to the OAIS for use in the construction or update of one or more Archival Information Packages (AIPs).

The SIP Creator is a web-based interface allowing the creation of E-ARK Submission Information Packages (SIPs). Files can be uploaded to the corresponding sections (content, documentation and metadata) of the container and a set of tasks allows transforming the SIP into the required format.

The SIP creator allows uploading files in the corresponding IP container sections.

Create SIP

SIP to AIP conversion

The AIP – as defined in the OAIS reference model – is an information package used to transmit and/or store archival objects within a digital repository. An AIP contains both, structural and descriptive metadata about the content, as well as the actual content itself.

The SIP to AIP conversion is a set of tasks that can be performed to convert an E-ARK SIP to an E-ARK AIP which both must comply with structural and metadata requirements defined by the E-ARK project.

Convert SIP to AIP

AIP to DIP conversion

The Dissemination Information Package (DIP) – as defined in the OAIS reference model – is the information package which is received by the consumer in response to a request for content.

The AIP to DIP conversion is a set of tasks that can be performed to convert one or several E-ARK AIPs to an E-ARK DIP.

Convert AIP to DIP

Figure 4: The earkweb user interface showing the four main areas SIP creator, SIP to AIP, AIP to DIP and Public search.

The common denominator of the "SIP creator", "SIP to AIP", and "AIP to DIP" areas is that they all offer information package transformation tasks. The transformation of information packages is implemented in the same way across all of the three information package transformation areas. The "SIP creator" and the "AIP to DIP" areas additionally provide some basic setup forms in order to collect information needed to initiate a new process. As shown in Figure 5, the "SIP creator" provides a form which allows for uploading individual files into the corresponding areas of the information package.

The interface for executing tasks is basically the same across all package transformation areas. The difference lies in the tasks they provide. Figure 6 shows the task execution interface of the "SIP to AIP" conversion. The pull-down select field shows tasks that are available in this area. Here, the available tasks are related to information packages which are converted from the initially submitted SIP to the AIP, which is finally transmitted to the long-term storage and/or uploaded into the distributed storage area for full-text indexing and access.

Figure 7 shows a search interface used in the "AIP to DIP" dialog that allows one to discoverer data in AIPs, select individual items, and generate DIPs.

5.4 Asynchronous and Parallel Package Processing

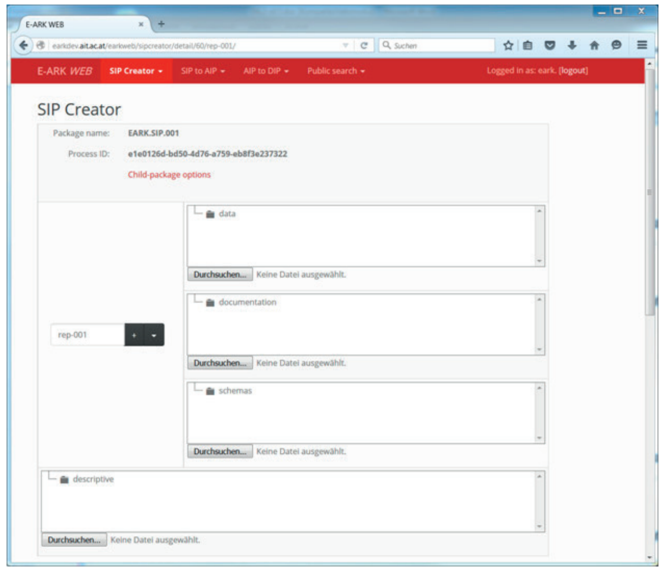
As mentioned in section 5.3, the transformation of information packages is implemented in the same way across all of the three information package transformation areas. In this section, we describe the task execution infrastructure used by the E-ARK Web project to enable the reliable and controlled execution of information package transformation tasks. Apart from the Python/Django-based user interface, E-ARK Web uses a backend for asynchronous and parallel task execution based on the Celery task execution system,

[6] https://github.com/eark-project/earkweb

[7] http://www.celeryproject.org

[8] https://www.python.org

[9] https://www.djangoproject.com


E-ARK WEB
earkdev.ait.ac.at/earkweb/sipcreator/detail/60/rep-001/
Suchen
E-ARK WEB
SIP Creator
SIP to AIP
AIP to DIP
Public search
Logged in as: eark. [logout]

SIP Creator

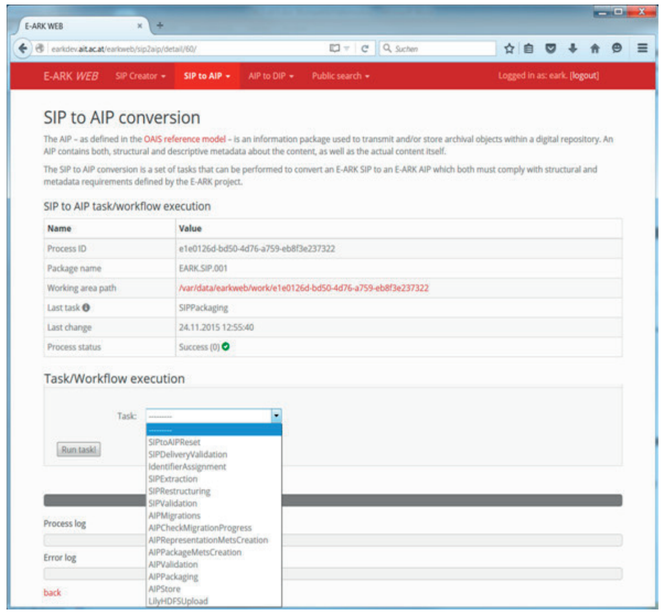
Package name: EARK.SIP.001

Process ID: e1e0126d-bd50-4d76-a759-eb8f3e237322

Child-package options

rep-001

Figure 5: User Interface of the SIP creator providing a form to select individual files and the corresponding location within the information package.


E-ARK WEB
earkdev.ait.ac.at/earkweb/sip2aip/detail/60/
Suchen
E-ARK WEB
SIP Creator
SIP to AIP
AIP to DIP
Public search
Logged in as: eark. [logout]

SIP to AIP conversion

The AIP – as defined in the OAIS reference model – is an information package used to transmit and/or store archival objects within a digital repository. An AIP contains both, structural and descriptive metadata about the content, as well as the actual content itself.

The SIP to AIP conversion is a set of tasks that can be performed to convert an E-ARK SIP to an E-ARK AIP which both must comply with structural and metadata requirements defined by the E-ARK project.

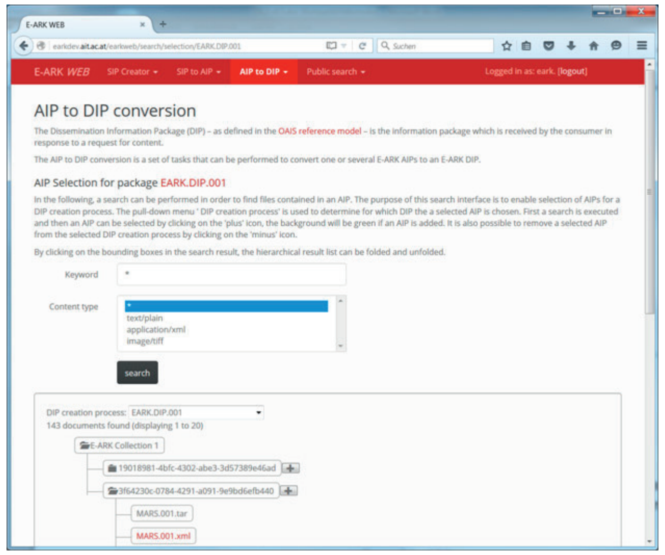
SIP to AIP task/workflow execution

Name	Value
Process ID	e1e0126d-bd50-4d76-a759-eb8f3e237322
Package name	EARK.SIP.001
Working area path	/var/data/earkweb/work/e1e0126d-bd50-4d76-a759-eb8f3e237322
Last task	SIPPackaging
Last change	24.11.2015 12:55:40
Process status	Success (0)

Task/Workflow execution

Task: ----------

Figure 6: User interface for selecting and starting an information package transformation. This screenshot shows the SIP to AIP conversion area.


E-ARK WEB
earkdev.ait.ac.at/earkweb/search/selection/EARK.DIP.001
Suchen
E-ARK WEB
SIP Creator
SIP to AIP
AIP to DIP
Public search
Logged in as: eark. [logout]

AIP to DIP conversion

The Dissemination Information Package (DIP) – as defined in the OAIS reference model – is the information package which is received by the consumer in response to a request for content.

The AIP to DIP conversion is a set of tasks that can be performed to convert one or several E-ARK AIPs to an E-ARK DIP.

AIP Selection for package EARK.DIP.001

In the following, a search can be performed in order to find files contained in an AIP. The purpose of this search interface is to enable selection of AIPs for a DIP creation process. The pull-down menu ' DIP creation process' is used to determine for which DIP the a selected AIP is chosen. First a search is executed and then an AIP can be selected by clicking on the 'plus' icon, the background will be green if an AIP is added. It is also possible to remove a selected AIP from the selected DIP creation process by clicking on the 'minus' icon.

By clicking on the bounding boxes in the search result, the hierarchical result list can be folded and unfolded.

Keyword: *

Content type: *

DIP creation process: EARK.DIP.001

143 documents found (displaying 1 to 20)

E-ARK Collection 1
19018981-4bfc-4302-abe3-3d57389e46ad
3f64230c-0784-4291-a091-9e9bd6efb440
MARS.001.tar
MARS.001.xml

Figure 7: Search interface in the AIP to DIP dialogue allowing a user to discover and select relevant data items from AIPs available in the Content Repository.

the MySQL[10] database, and the RabbitMQ[11] message broker software.

Whenever a task is initiated using the E-ARK Web task execution interface, the RabbitMQ message broker receives a message which is subsequently consumed by the Celery task execution engine. Tasks can be assigned to workers which are configured in the Celery backend. The workers share the same storage area and the result of the package transformation is stored in the information package's working directory based on files.

As the actual status of the transformation process is persisted during the task execution it is not required to interrupt the processing chain for every executed task in order to update status information in the database. Based on the results stored in the working directory, the status of an information package transformation can be updated with a single operation when the transformation process has finished. This strategy increases the processing efficiency, which is critical when large volumes of data are processed, and helps avoiding bottlenecks caused by a large number of parallel database connections. Another advantage of this approach is that by design it is possible to reconstruct the databases, tracking the status of the processed information package, based on the information contained in the working directories. Particular importance was given to the principle of avoiding to instantly record digital object related processing information in the database as this may entail the risk of significantly increasing the processing time for very large information packages.

The decision to use either synchronous or asynchronous task execution for a specific task depends on the type of task and also the kind of data the information package contains. A task which itself initiates an unknown number of sub-tasks, can lead to a long task runtime, possibly beyond the defined timeout limit. An example would be a set of file format migration sub-tasks which are triggered for specific file types, e.g. each PDF file contained in an information package is converted to PDF/A. These cases can be implemented using a master task that starts an unknown number of sub-tasks and records the amount of migrations to be performed. This task is followed by a verification task which can be executed manually or automatically to report the current status of task executions. This way, it is possible to control that subsequent executions are not started before all sub-tasks were executed successfully, and that all the (possibly long-running) processes are decoupled from each other. The upload of an AIP into the Hadoop infrastructure has been implemented as a synchronous task. The live progress of the upload process is shown directly in the user interface. However, if for cases where AIPs tend to be very large – where "large" is to be seen in the context of available bandwidth and read/write throughput – it is easily possible to change this task execution into an asynchronous task

[10] https://www.mysql.com
[11] https://www.rabbitmq.com

5.5 Task and Workflow Definition

With respect to software design, a major goal was to foster flexibility, modularity, and extensibility of the task execution base class. Tasks are implemented in one single Python script and only contain the code that is necessary for the concrete task implementation. The intention is to keep the actual task implementation slim and offload extensive functionality into an earkcore Python module[12] which can be made available to the Celery workers.

The E-ARK Web project defines a workflow model on top of the task execution layer. The "state" of an information package, as described earlier, is defined by storing the "last executed task" together with the success/failure of the execution. Tasks provide interface definitions (like for example "allowed inputs") which provide the basis for workflow composition. Using this information together with the current execution status, the workflow engine can control if a task is allowed to be performed on a specific information package.

New tasks can be easily added to the system by supplying a new task class implementation based on a Python script. The new task is available in the system as soon as the Celery workers are re-initialized. The configuration of the task is handled directly within the task implementation based on code annotations. Information to verify workflow composition is immediately available through the task description and does not require any additional configuration files. As the descriptive information is used to initialize the task configuration information in the database, it can be also dynamically adapted in the database, if required.

6. EXPERIMENTAL EVALUATION

6.1 Hardware Environment

The Lily/Hadoop deployment on the development cluster at AIT is shown in figure 8. The cluster comprises seven physical machines which are structured into a master and six physical slave nodes. Each node on the cluster provides 6 CPU cores (12 threads using Intel HT), 16GB RAM and 16TB SATA (hotplug) of storage. Each cluster node is equipped with two network interfaces allowing us to attach a node to two network infrastructures. The cluster is connected to the internal network allowing us to directly access each node from desktop/working environments. The second private network is used for managing the cluster. For example, new cluster nodes can be automatically booted and configured using the PXE pre-boot execution environment together with a private Fully Automated Install (FAI) server[13].

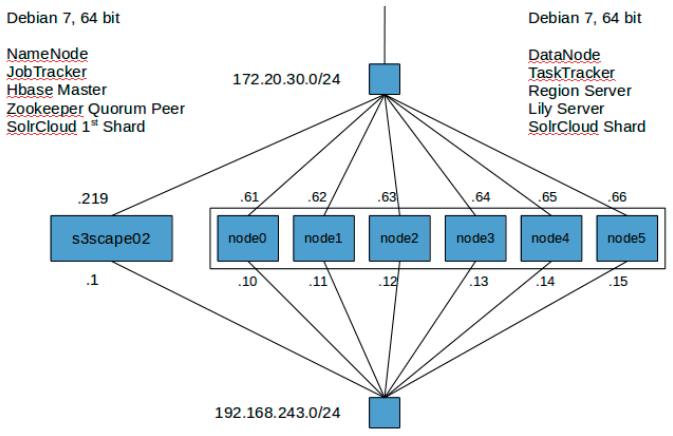


Figure 8: Hardware cluster at AIT used to host a Lily repository on top of Hadoop, HDFS and HBase.

6.2 Data Set

The govdocs1 corpus [8] is a set of about 1 million files that are freely available for research. This corpus provides a test data set for performing experiments using different types of typical office data files from a variety of sources. The documents were originally obtained randomly from web servers in the .gov domain. Due to the volume of collected files and the variety of data types available in this corpus, we have chosen to perform a document discovery over the entire corpus as a simple use case for evaluating for the E-ARK integrated prototype.

Here, it is important to note that the integrated prototype is designed for the ingestion of information packages as described by the OAIS model. E-ARK is developing a general model along with as set of specifications and tools for handling SIP, AIP, and DIP packages, which are being included with the integrated prototype's package creation infrastructure. AIPs are typically created as structured tar-files containing data and metadata as described by the E-ARK AIP format[14]. The repository provided by the integrated prototype is designed to maintain the structure of the ingested information packages (by encoding file locations within the record identifier) — allowing users to browse and search single packages if desired — but in general provides search and access across information packages on a per-file basis. For the experimental evaluation we have ingested the govdocs1 corpus in the form of 1000 tar files, each containing 1000 documents, which results in 1000 packages available in the integrated prototype's repository, and 1 million files that are full-text indexed, and that can be individually identified by an URL and accessed via the REST API.

[12] https://github.com/eark-project/earkweb/tree/master/earkcore

[13] http://fai-project.org
[14] http://www.eark-project.com/resources/project-deliverables/53-d43earkaipspec-1

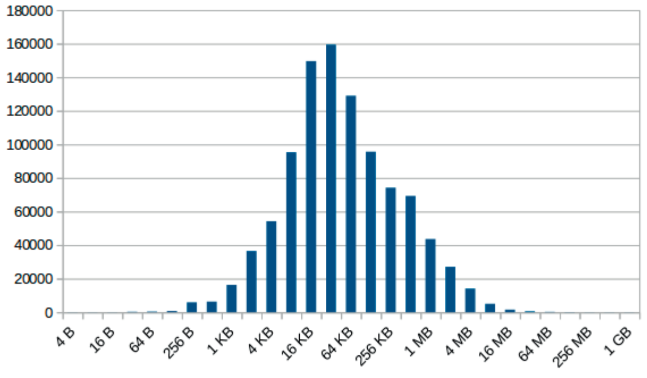


Figure 9: Govdocs1 file size distribution

Table 2: I/O performance benchmarking

TestDFSIO	*write*	*read*
Number of files	10	10
Total MBytes processed	10000.0	10000.0
Throughput mb/sec	17.80604586481	49.9852543499
Average IO rate mb/sec	17.93034553527	52.1930541992
Test exec time sec	72.661	36.593

Once all packages are ingested, documents can be found using the search server API. Queries might include a full text search string, filtering based on metadata (like MIME-type), or restrict the search results on certain packages. Using facets in a search query allows one to easily derive general statistics about a search result. Figure 9 illustrates the result of a faceted search query which groups all files of the ingested govdocs1 corpus based on file-sizes. Most of the files fall in the range between 1KB and 16MB and only a few small files with size values starting from 7 bytes and 4 text files over 1.5 gigabytes exist. An overview of the MIME types available in the corpus is described by [17, p. 15]. We will show as part of this evaluation how to retrieve this kind of information from the system once the collection has been successfully ingested.

6.3 Cluster I/O Benchmarking

To provide indicative benchmarks, we executed the Hadoop cluster I/O performance benchmarking test "TestDFSIO" as described by [15] which is a read and write benchmarking test for the Hadoop Distributed File System (HDFS). TestDFSIO is designed in such a way that it uses 1 map task per file. This is similar to the file ingest component of the integrated prototype where each package (available as a TAR file) is processed by one task. The default test method of TestDFSIO is to generate 10 output files of 1GB size for a total of 10GB in the write test which are subsequently read by the "read" test. The results of this test are as presented in table 2.

6.4 Evaluation Results

The purpose of this evaluation is to give an approximate insight on the performance of the E-ARK integrated prototype. Due to the complexity of the system set-up and the numerous configuration options, the presented results should only provide an indication of the achieved cluster performance rather than provide strict benchmarking results.

Table 3: The integrated prototype automatically triggers a MapReduce job when ingesting data into the repository. The table shows the results reported by the Hadoop MapReduce execution environment after the govdocs 1 corpus has been ingested as a set of 1000 tar-files.

Hadoop Job	*File Ingest Mapper*
Number of map tasks	1000
Map input records	984951
Map output records	354
Job finished in	1hrs, 47mins, 51sec

Table 4: Parameters of a faceted query that orders the search results by the number of by MIME-types.

Query parameter	*Value*
facet	on
q	*:*
facet.field	contentTypeFixed
rows	0

We defined a threshold for the file ingest workflow (executed as a map task) to process a maximum file size of 50 Megabytes. The Govdocs1 corpus contains 354 files exceeding this limit. These files sum up to a total size of about 42 Gigabytes and were ingested separately. The pre-configured file limitation is an implementation detail which has been set for practical reasons. In case it is required to automatically ingest files of large sizes, this can be handled as well. While Lily stores small files in HBase for efficient random access, large files are stored directly in HDFS. There is no file size limitation regarding the ingest or storage of files in the repository. The basic test results of the Hadoop job performing the ingest workflow are shown in table 3.

The number of 1000 map tasks corresponds to the 1000 TAR packages of the Govdocs1 corpus which were defined as the input of the Hadoop job. The 984951 input records are the individual files which were found in the TAR packages. The map task performs the ingest of files into Lily and outputs only those files which had been skipped due to their file size, as described earlier. The set of 1000 processed tar-files sums up to a total of 467GB and the total wall time for the ingest process amounts to 1 hour and 47minutes.

The files contained in the ingested tar-files are searchable using the Solr search interface. Part of the job execution was to run text extraction and MIME-Type detection using Apache Tika and to store this information in the index, therefore it is now possible to run a single faceted Solr query to get basic MIME-Type statistics with the parameters specified in table 4, where the field "contentTypeFixed" is the field of type "string" defined in the schema of the collection which holds the MIME-type of the corresponding file item. This allows us, for example, to get an overview about the ten most frequent MIME types in the collection as presented in figure 10.

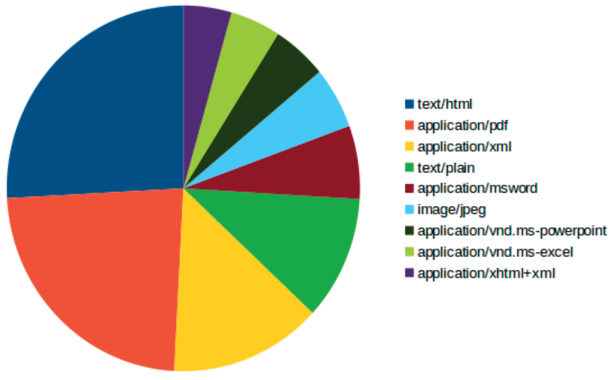


Figure 10: The ten most frequent MIME types (resulting from a Solr facet query)

7. ADVANCED DATA ACCESS SCENARIOS

As part of the ingest workflow, the integrated prototype adds, besides full-text information, various (often content and use-case specific) metadata elements to the the repository record in order to enhance the capabilities of the derived search index, as briefly demonstrated in the previous section. Ongoing experimental work is dealing with utilizing additional data mining strategies to further enhance the integrated prototype's search functionality. Two currently evaluated text mining strategies are Classification and Named Entity Recognition, as explained below. The goal is to add useful information to the full-text index, such as discovered text categories. Besides selecting appropriate algorithms and designing useful features, it is a challenge to run such data mining algorithms on very large data volumes. Initial experiments have been performed using ToMar [20], a MapReduce application for efficiently running 3rd party tools on a Hadoop-based cluster, developed in the context of the SCAPE project.

7.1 Text Classification

The classification of text allows one to make assumptions on the contents of files, based on a previously trained model. We are planning to use this technique to extend the search interface provided by the E-ARK Web user interface through adding a field for selecting automatically recognized text categories. This way it is possible to search for documents which relate to a specific topic combined with the existing query and filtering options provided by the search server. The number of search-able topics depends on the previously trained classifier process, and therefore include an assumption on which topics could be of interest for the user. As a toolkit for implementing text classification, we have utilized the *scikit-learn* [16] Python framework.

7.2 Named Entity Recognition

An additional goal was to identify locations, persons, or other terms of interest as so called Named Entities. In initial tests the the Stanford Named Entity Recognizer[7] has been utilized to extract entities from text documents. Entities that were classified as locations were, in an additional step, geo-referenced using the Nominatim database [4]. As a result, an XML file containing a list of found locations together with their corresponding coordinates was generated for each analyzed document. The intention behind this work is to incorporate new ways of making archived content accessible to the user. Initial experiments dealt with visualizing the geographical focus of identified topics over time using the graphical map annotation tool Peripleo [21].

8. RELATED WORK

Warcbase [13] uses HBase as the core technology to provide a scalable and responsive infrastructure for web archiving. The environment makes use of the random access capabilities of HBase to build an open-source platform for managing raw content as well as metadata and extracted knowledge. Additionally, Warcbase provides exploration, discovery, and interactive visualization tools that allow users to explore archived content.

The Internet Memory Foundation has built a distributed infrastructure for Web archiving, data management, and preservation using Apache Hadoop as one of the core technologies [14]. Their focus is on scalability issues in terms of crawling, indexing, preserving and accessing content.

RODA is an open source digital repository which delivers functionality for all the main units of the OAIS reference model [6]. RODA provides distributed processing and the execution of digital preservation actions (e.g. migration) on a Hadoop cluster.

The European project SCAPE (Scalable Preservation Environments) addressed the preservation of very large data sets found in digital repositories, scientific facility services, and web archives as one of the main use cases [18]. SCAPE has build on top of a Hadoop-based infrastructure for defining and carrying out preservation workflows. Additionally the project investigated the integration of an Hadoop-based infrastructure with the Fedora Commons repository systems [11].

9. CONCLUSIONS

In this paper we presented a prototype infrastructure for the scalable archiving of information packages developed in the E-ARK project. The system is implemented using a set of open source technologies including for example Apache Hadoop, the Lily Project, and Apache SolR. As we have outlined, there are a number of related projects, mostly in the Web archiving domain, which are using a similar technology stack for scalable e-archiving. The system presented in this paper is however targeting the archival community and specifically designed to support OAIS-based concepts. The Integrated Platform Reference Implementation Prototype has been developed to handle the creation, ingestion, and access of E-ARK information packages using an environment that scales from a single host to a cluster deployment. The system can be deployed as a stand-alone environment but also next to existing archiving systems in order to enhance available services, like for example finding aids (using the full text index) and order management (using the content repository).

Here, we have provided a brief overview of the system architecture and the employed technologies. We have also described the ingest workflow in more detail and explained how the individual components are employed and how they are related to each other. As an evaluation of the approach, we have ingested and indexed the entire Govdocs1 corpus consisting of nearly 1 million documents with a total size of about 467 Gigabytes in less then 2 hours, making the text content discoverable in and across information packages based on full-text as well as metadata-based queries using a powerful search server. The used repository provides instant access at the granularity of single files which can be viewed and/or packaged for dissemination using the provided E-ARK Web access components. The paper reports also on future directions to further improve the search capabilities of the system by employing data mining algorithms.

10. ACKNOWLEDGMENTS

Work presented in this paper is primarily supported by European Community's Seventh Framework Program through the E-ARK project under grant agreement number 620998.

11. REFERENCES

[1] V. Borkar, M. J. Carey, and C. Li. Inside "big data management": Ogres, onions, or parfaits? In *Proceedings of the 15th International Conference on Extending Database Technology*, EDBT '12, pages 3–14, New York, NY, USA, 2012. ACM.

[2] P. Caplan and R. S. Guenther. Practical preservation: the premis experience. *Library Trends*, 54(1):111–124, 2006.

[3] CCSDS. *Reference Model for an Open Archival Information System (OAIS)*. CCSDS - Consultative Committee for Space Data Systems, January 2012. Version 2 of the OAIS which was published in June 2012 by CCSDS as "magenta book" (ISO 14721:2012).

[4] K. Clemens. Geocoding with openstreetmap data. *GEOProcessing 2015*, page 10, 2015.

[5] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. In *OSDI'04: Proceedings of the 6th conference on operating systems design and implementation*. USENIX Association, 2004.

[6] L. Faria, M. Ferreira, R. Castro, F. Barbedo, C. Henriques, L. Corujo, and J. C. Ramalho. Roda - a service-oriented repository to preserve authentic digital objects. In *Proceedings of the 4th International Conference on Open Repositories*. Georgia Institute of Technology, 2009.

[7] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *In ACL*, pages 363–370, 2005.

[8] S. Garfinkel, P. Farrell, V. Roussev, and G. Dinolt. Bringing science to digital forensics with standardized forensic corpora. *digital investigation*, 6:S2–S11, 2009.

[9] A. J. Hey, S. Tansley, K. M. Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA, 2009.

[10] A. J. Hey and A. E. Trefethen. The data deluge: An e-science perspective. 2003.

[11] B. A. Jurik, A. A. Blekinge, R. B. Ferneke-Nielsen, and P. Møldrup-Dalum. Bridging the gap between real world repositories and scalable preservation environments. *International Journal on Digital Libraries*, 16(3):267–282, 2015.

[12] R. King, R. Schmidt, C. Becker, and S. Schlarb. Scape: Big data meets digital preservation. *ERCIM News*, 89:30–31, 2012.

[13] J. Lin, M. Gholami, and J. Rao. Infrastructure for supporting exploration and discovery in web archives. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 851–856, New York, NY, USA, 2014. ACM.

[14] L. Medjkoune, S. Barton, F. Carpentier, J. Masanès, , and R. Pop. Building scalable web archives. In *Archiving Conference, Archiving 2014 Final Program and Proceedings*, number 1, pages 138–143, 2014.

[15] M. Noll. Benchmarking and stress testing an hadoop cluster with terasort, testdfsio & co. http://www.michael-noll.com/blog/2011/04/09/benchmarking-and-stress-testing-an-hadoop-cluster-with-terasort-testdfsio-nnbench-mrbench, 4 2011.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[17] M. Radtisch, P. May, A. A. Blekinge, and P. Moldrup-Dalum. D9.1 characterisation technology, release 1 and release report. http://www.scape-project.eu/wp-content/uploads/2012/03/SCAPE_D9.1_SB_v1.0.pdf, 04 2012.

[18] S. Schlarb, P. Cliff, P. May, W. Palmer, M. Hahn, R. Huber-Moerk, A. Schindler, R. Schmidt, and J. van der Knijff. Quality assured image file format migration in large digital object repositories. In J. Borbinha, M. Nelson, , and S. Knight, editors, *iPRES 2013: 10th International Conference on Preservation of Digital Objects, 2-6 September 2013*, pages 9–16, Lisbon, Portugal, September 2013. Lisbon Technical University (IST).

[19] R. Schmidt. An architectural overview of the scape preservation platform. In *9th International Conference on Preservation of Digital Objects*, pages 85–55. Citeseer, 2012.

[20] R. Schmidt, M. Rella, and S. Schlarb. Constructing scalable data-flows on hadoop with legacy components. In *11th IEEE International Conference on e-Science, e-Science 2015, Munich, Germany, August 31 - September 4, 2015*, page 283, 2015.

[21] R. Simon, E. Barker, L. Isaksen, and P. de Soto Cañamares. Linking early geospatial documents, one place at a time: Annotation of geographic documents with recogito. *e-Perimetron*, 10(2):49–59, 2015.

Securing Trustworthy Digital Repositories

Devan Ray Donaldson
School of Informatics and Computing
Indiana University
1320 E. 10th St., Wells Library 019
Bloomington, IN 47405-3907
+1-812-855-9723
drdonald@indiana.edu

Raquel Hill
School of Informatics and Computing
Indiana University
230E Lindley Hall
Bloomington, IN 47406
+1-812-856-5807
ralhill@indiana.edu

Heidi Dowding
Library Technologies
Indiana University Libraries
1320 E. 10th St., Wells Library W501
Bloomington, IN 47405-3907
+1-812-856-5295
heidowdi@indiana.edu

Christian Keitel
Landesarchiv Baden-Württemberg
Eugenstraße 7
D-70182 Stuttgart
+49 0711-212-4276
christian.keitel@la-bw.de

ABSTRACT

Security is critical to repository trustworthiness. Recent international standards for Trustworthy Digital Repositories (TDRs) all specify some sort of security criteria that are necessary to adhere to in order to attain TDR status. However, little is known about how those who are responsible for addressing these criteria actually regard the concept of security. This study centers on digital repository staff members' perceptions of security, including their perceptions of security criteria in standards for TDRs. This paper discusses findings from surveys and semi-structured interviews with staff from repositories that have recently acquired the nestor seal of approval. We found that participants considered the principles of confidentiality, integrity, and availability as relevant to their notions of security. We also found that participants considered the security criteria required to acquire the nestor seal of approval as both sufficient and appropriate for addressing their repositories' needs. Implications for better understanding the security of digital repositories are discussed as well as directions for future research.

Keywords

Security; Trustworthy Digital Repositories; Repository Staff Perceptions.

1. INTRODUCTION

Unarguably, security is part of what is necessary for a digital repository to be trustworthy. Evidence of the importance of security can be seen by examining criteria pertaining to security in recent standards for Trustworthy Digital Repositories (TDRs). For example, these criteria specify that staff identify sections of their repositories that are worthy of protection, analyze potential threats and perform risk assessment [4, 5, 9]. While security criteria in standards for TDRs seem relatively straightforward, little is known about actual staff members' perceptions of these security criteria. For example, staff may consider the criteria relatively easy to address, or they may consider the criteria rather challenging to address. Staff also may consider their repositories more secure as a result of adhering to these criteria or they may not. Digital repository staff members have a direct impact on the security of TDRs. They make decisions and implement policies that can result either in increased security or compromises to security. For these reasons it is critically important to better understand how digital repository staff members think about security.

The purpose of this study is to understand digital repository staff members' perceptions of security for TDRs. The remainder of this paper is as follows. First, we explore scholarship on security in the digital preservation and computer science literatures. Second, the methodology section describes the sample of participants and explains why they were selected. The methodology section also describes data collection and analysis techniques. Third, the findings are reported. The paper concludes with a discussion and explication of implications of the study and recommends directions for future research.

2. SCHOLARSHIP ON SECURITY

2.1 Security in the Digital Preservation Literature

Security refers to "the practice of defending information from unauthorized access, use, disclosure, disruption, modification, perusal, inspection, recording or destruction" [16, p. 224]. The best place to understand the phenomenon of security within the field of digital preservation is to examine recent standards for TDRs. They represent a consensus among key members of the digital preservation community on what constitutes best practice. They include specific criteria pertaining to security as part of attaining formal "trustworthy" status for digital repositories. For example, criterion C34 in DIN 31644 requires organizations and their infrastructures to protect their digital repositories and their contents [4, 11]. In particular, criterion C34 requires staff at organizations to protect the integrity of digital repositories and their content. To accomplish this, nestor certification against criterion C34 recommends that staff identify sections of the archive that are worthy of protection, analyze potential threats to the archive, and perform risk assessment "of the damage scenarios [to] ultimately result in a consistent security system" [11, p. 40]. For example, according to the explanatory notes on the nestor seal for TDRs, criterion C34 asks staff at organizations to identify which of three types of damage scenarios they perceive as a particular threat to information preserved by digital repositories: 1) malicious actions, 2) human error, or 3) technical failure. The explanatory notes also ask staff to consider the likelihood of each damage scenario, the seriousness of each scenario as well as what level of residual risk is acceptable. Furthermore, they ask staff about what measures they are taking to counter these risks as well as how they plan to implement their risk analysis and planned countermeasures into their security systems. Finally, these notes ask staff about their plans to test and further develop their security systems.

Similarly to DIN 31644, ISO 16363 includes a section on security entitled “Security Risk Management” [9]. This section outlines security criteria for TDRs. According to ISO 16363, staff seeking “trustworthy” status for their digital repositories must maintain “a systematic analysis of security risk factors associated with data, systems, personnel, and physical plant” [9, p. 76]. A TDR must also:

- Implement controls to address defined security risks,
- Have delineated roles, responsibilities, and authorizations related to implementing changes within the system, and
- Have suitable written disaster preparedness and recovery plans.

ISO 16363 also describes three additional security concerns that could arise during audit. First, the auditor could be a false auditor or have malicious intent. Second, confidential information could be lost as a result of performing the audit, which could compromise the system. Third, damage to the repository system could occur while transferring information during audit. To guard against these security threats, recommendations in ISO 16363 include:

- Relying on repositories’ identification and authorization systems,
- Relying on the security systems of auditors and settling on information transfer agreements between repositories and auditors, and
- Relying on repositories’ security and safety systems.

Both DIN 31644 and ISO 16363’s security requirements draw upon an earlier standard for TDRs: Digital Repository Audit Method Based on Risk Assessment known as DRAMBORA [5]. For example, ISO 16363 recommends that digital repository staff members use DRAMBORA as a tool for performing risk assessments. Similarly, DIN 31644 recommends that digital repository staff members use DRAMBORA to help identify the sections of the archive which are worthy of protection, analyze any potential threats to the specific archive, and perform risk assessments of possible damage scenarios.

The DRAMBORA methodology consists of six steps. First, digital repository staff members should identify their objectives. DRAMBORA includes a list of examples of objectives for digital repository staff members to choose from. Second, digital repository staff members should identify the activities that are necessary to achieve their objectives and assets, including human resources and technological solutions, that are central to achieving repositories’ objectives. Third, digital repository staff members should align risks to their activities and assets. This step requires digital repository staff members to document the specific risks associated with each identified activity and asset. Here a single risk may associate with multiple activities, or vice versa. Fourth, digital repository staff members should assess, avoid, and treat risks by characterizing each risk’s “probability, impact, owner, and the mechanisms or proposed mechanisms by which it can be avoided or treated” [5, p. 39]. Fifth, digital repository staff members should self-audit their repositories to determine what threats are most likely to occur and identify areas where improvement is required. Sixth, digital repository staff members should complete a risk register listing all identified risks and the results of their analysis and evaluation. Also known as a risk log, it should include information about the status of each risk and include details that can aid digital repository staff members in tracking and monitoring risks.

Taken together, standards for TDRs underscore the importance of security and provide relatively similar recommendations to digital repository staff members about how to address security. However, the security criteria themselves do nothing to illuminate actual digital repository staff members’ perspectives on security or their perceptions of the said security criteria.

2.2 Security in the Computer Science Literature

Relevant to a discussion on security in the digital preservation literature is discussion of security in the computer science literature. In digital preservation, the primary focus is on the security of digital repositories and their content. On the other hand, in the field of computer science security is more encompassing, including a broad range of computing infrastructures, not just digital repositories. Computer science also has a longer, more established body of literature on security, including definitions and metrics for the concept.

Computer scientists who specialize in security research have reached a consensus that computer security consists of at least three main principles: confidentiality, integrity, and availability. Confidentiality refers to concealment of information or resources, integrity refers to the trustworthiness of data or resources, and availability refers to the ability to use the information or resource desired [1]. While security researchers seem to agree on these three principles of security, others have proposed additional security elements. For example, some researchers have recommended including the concept of accountability, “the security goal that generates the requirement for actions of an entity to be traced uniquely to that entity,” in defining trustworthiness [17, p. A-1]. As another example, OECD guidelines proposed nine security principles: awareness, responsibility, response, ethics, democracy, risk assessment, security design and implementation, security management, and reassessment [12]. Stoneburner, Hayden, and Feringa [17] proposed thirty-three principles related to having a security foundation, risk, ease of use, increasing resilience, reducing vulnerabilities, and designing with the network in mind. Parker [13] extended the classic Confidentiality-Integrity-Availability (CIA) triad by adding three elements: possession, authenticity, and utility. After a thorough review of the literature, Cherdantseva and Hilton [3] proposed extending the CIA triad to an Information Assurance and Security (IAS) octave consisting of: confidentiality, accountability, auditability, authenticity/trustworthiness, non-repudiation, and privacy. It is important to note that Cherdantseva and Hilton had IAS academics and experts evaluate the IAS octave. According to Cherdantseva and Hilton, the IAS octave is part of a larger, all encompassing reference model of information assurance and security. Although alternative models of security exist, all seem to incorporate confidentiality, integrity, and availability at their core.

In addition to multiple definitions of security, the literature on security in computer science also offers some security metrics. For example, these metrics can provide assessment of security properties, measurement of adherence to secure coding standards, monitoring and reporting of security status, and gauge the effectiveness of various security controls [7, 10]. Although some security metrics exist, researchers acknowledge that security is actually quite difficult to measure. Pfleeger and Cunningham [14] list nine reasons why security is hard to measure:

- We can’t test all security requirements,
- Environment, abstraction, and context affect security,
- Measurement and security interact,
- No system stands alone,
- Security is multidimensional, emergent and irreducible,
- The adversary changes the environment,
- Measurement is both an expectation and an organizational objective,
- We’re overoptimistic, and
- We perceive gain differently from loss.

Common to both computer security and security for digital repositories is threat modeling. During the threat modeling process, assets are identified; threats against the assets are enumerated; the likelihood and damage of threats are quantified; and mechanisms for mitigating threats are proposed [2, 5, 8, 9, 11, 15].

While some components of the threat modeling process are qualitative, quantifying the risk of threats enables system administrators to rank the order in which threats should be addressed. Within the computer science literature, various approaches have been proposed for characterizing and quantifying the risk of threats, including calculating risk as the product of the damage potential and the likelihood of occurrence, *Risk = Criticality * Likelihood of Occurrence* [8]. Dread, an approach proposed by Microsoft, calculates risk across several categories, including: Damage potential, Reproducibility, Exploitability, Affected users, and Discoverability [8]. Using Dread, a threat is rated on a scale from 1 to 10 for each category, with the resulting risk being the average of all ratings. Butler and Fischbeck [2] propose a multiple attribute threat index (TI) for assessing the risk of a threat. TI captures the relative importance of each type of threat [2], where $TI_a = Freq_a * (\sum_{j=attributes} W_j * X_{aj})$, and W*j* is the attribute weight and X*aj* is the most likely outcome value for the threat.

While quantifying risks will enable us to capture an organization’s security requirements, Pfleeger [15] advises that we should avoid false precision by doing the following:

- Base the probability distribution of a threat/attack occurring on historical data, not just on expert judgment;
- Since "both scientists and lay people may underestimate the error and unreliability in small samples of data, particularly when the results are consistent with preconceived, emotion-based beliefs”, we are to be mindful of the size of our experiments and the scalability of our results.

While measuring security is difficult, and few security metrics of any kind exist, metrics for understanding perceptions of security are particularly scant.

Taken together, the literature on security in digital preservation and computer science stress the importance of security, while also leaving several open research questions. This study focuses on four of them:

1. How do digital repository staff members think about the security of Trustworthy Digital Repositories?
2. What are digital repository staff members’ attitudes toward security criteria in standards for Trustworthy Digital Repositories?
3. How relevant are security principles that have been established in the computer science domain to digital repository staff members’ concept of security?
4. Is it possible to develop a survey that could serve as a tool for measuring digital repository staff members’ perceptions of security for Trustworthy Digital Repositories?

3. METHODS

To address the research questions, we conducted interviews with digital repository staff members at organizations whose repositories have attained formal, third-party trustworthy status. We also administered surveys to those individuals. The purpose of using these data collection methods was to understand how the participants thought about security and to assess measurement of the concept. While various standards for trustworthy digital repositories exist [4, 6, 9], at present, DIN 31644 is the only standard that: 1) has been formally recognized by a standards-granting body, and 2) has organizations whose repositories have been formally certified by third parties. Thus, we decided to include in our study only digital repository staff members whose repositories have recently acquired nestor seals of approval, signifying formal, third-party certification by the DIN 31644 standard. To date, two organizations have successfully acquired nestor seals of approval. During April 2016, we recruited participants at these institutions via email, asking them to participate in our interviews and take our survey.

3.1 Interviews

During semi-structured interviews, participants discussed their definitions of security and trustworthiness. They also discussed their views on the relationship between the trustworthiness of digital repositories and their security. Afterwards, participants discussed security criteria in DIN 31644 (e.g., criterion C34), including how easy or difficult they thought it was to address the criteria, how prepared they felt to address the criteria, how they approached addressing the criteria, whether they thought the criteria were sufficient, and what, if any, additional criteria they would recommend. Participants also discussed the extent to which they thought their repositories were more secure as a result of adhering to these criteria. Appendix A includes the interview protocol. The interviews lasted approximately 30 minutes and took place on Skype.

All interviews were audio-recorded and transcribed. Afterwards, transcripts were coded using NVivo – a qualitative data analysis software tool. Prior to analyzing the transcripts, we developed a codebook based primarily on the three main dimensions of security established in the computer science literature: confidentiality, integrity, and availability. Specifically, two members of the research team coded the transcripts looking for any statements participants made that corresponded to the concepts of confidentiality, integrity, and availability. We then calculated a table enumerating the frequencies with which participants mentioned each concept in relation to their perceptions. Finally, we calculated inter-rater reliability using Cohen's kappa, achieving a score of 0.79.

3.2 Surveys

In developing our survey, we examined the literature on security in digital preservation and computer science, including research on security metrics. We did not find an existing instrument to measure the security perceptions of computing infrastructures by those who are responsible for managing and securing said infrastructure. Consequently, we derived items for our survey from definitions and explanations of confidentiality, integrity, and availability in Bishop [1], a foundational text on computer security.

We asked the same individuals that we interviewed to take our survey. The survey consisted of 19 items: 4 pertaining to confidentiality, 11 pertaining to integrity, and 4 pertaining to availability. The survey included a 5-point, likert-type scale ranging from "Strongly disagree" to "Strongly agree" with one additional option: "Not applicable." Appendix B includes the survey instrument. The items were randomized to mitigate order effects.

To analyze the survey data, we calculated descriptive statistics, including participants' mean scores on the items that pertained to confidentiality, integrity, and availability. We also performed the Kruskal-Wallis H test to identify whether there were any statistically significant differences in participants' attitudes toward the confidentiality, integrity, and availability principles.

4. FINDINGS

The findings are organized based on the methods we used to collect the data. After discussing participant characteristics, we discuss findings from the interviews. Next, we discuss findings from the surveys.

4.1 Participant Characteristics

Two people participated in this study, one from each organization that successfully acquired the nestor seal of approval. Both participants held senior positions in the organizations where they worked. Their responsibilities included overseeing teams involved in national and international digital preservation projects and initiatives as well as policy and services development within their organizations. Participants reported working approximately five to nine years on digital repositories at their current organizations. Both participants reported having involvement in the development of standards for digital repositories.

4.2 Interview Findings

Participants shared their views on the concept of security for digital repositories. Specifically, they viewed security as a prerequisite for trustworthiness. They saw security as making sure that repositories act as they are supposed to with no intended or unintended interruptions.

Participants also shared their views on criterion C34 and its explanatory notes. They thought that criterion C34 itself was a bit general, but the explanatory notes for C34 were a helpful complement, providing guidance on how to successfully address the criterion within their repositories. Despite the fact that participants found it difficult to address criterion C34, they felt prepared to address it based on the security measures they had in place prior to audit (e.g., redundant storage, protection against data manipulation, and implementation of national IT standards). While participants did not consider their repositories more or less secure as a result of addressing the explanatory notes for criterion C34, they thought their documentation for what they do to secure systems improved. When asked whether the explanatory notes for criterion C34 set the bar for security too high, too low, or just right, participants stated that addressing the explanatory notes sets the bar just right, suggesting that they considered the security requirements for nestor certification as reasonable, appropriate, and sufficient for securing their repositories.

Analysis of interview data against the confidentiality, integrity, and availability security principles established in computer science revealed that participants provided statements pertaining to the concept of integrity most frequently, followed by availability and confidentiality. Table 1 lists the frequency with which participants provided statements pertaining to each concept. When participants mentioned integrity, they referred to protecting their data from any threats, including manipulation. Participants mentioned the importance of confidentiality and availability because both are included in the nestor definition of security—a definition which they reported as being important to their work. They did not, however, elaborate on what either of the concepts meant to them in their practice.

Table 1. Frequency Participants Mentioned Security Concepts

Security Concepts	Frequency
Confidentiality	2
Integrity	10
Availability	2

4.3 Survey Findings

To complement the interview data and get a better sense of the relevance of security principles to the participants, we administered surveys to them. The surveys asked questions about participants' views on aspects of confidentiality, integrity, and availability.

Table 2 lists the mean scores of participants' responses for the questions pertaining to each security principle. Comparing the mean scores of participants' responses to the survey questions reveals that participants are most concerned with integrity, followed by availability and confidentiality.

Table 2. Mean Scores for Security Concepts

Security Concepts	Mean Scores
Confidentiality	3.38
Integrity	4.55
Availability	3.75

A Kruskal-Wallis H test showed that there was a statistically significant difference in participants' ratings of security survey items based on the different principles the items referred to, $X^2(2) = 7.82$, $p = .02$, with a mean rank security score of 13.75 for confidentiality, 23.50 for integrity, and 14.25 for availability. These results suggest that participants had stronger attitudes about integrity relative to their attitudes about availability and confidentiality.

5. DISCUSSION

Results underscore the importance of security to the digital repository staff members who participated in this study. Participants mentioned the three security principles of confidentiality, integrity, and availability during the interviews. Participants also rated survey items pertaining to those three principles highly, suggesting that they are relevant to their views on securing digital repositories.

Although participants mentioned the three security principles of confidentiality, integrity, and availability during the interviews, and rated survey items pertaining to them highly, results of this study provide more empirical support for some principles of security than others. For example, participants provided more statements related to integrity than availability and confidentiality. As another example, participants rated survey items pertaining to integrity higher than survey items pertaining to availability and confidentiality. The fact that the interview data and survey data triangulate with respect to more emphasis on integrity relative to availability and confidentiality is interesting and needs to be looked at more in depth in future research. The main questions that we need to understand going forward are: Why is integrity more salient to digital repository staff members? And what might this mean for research and practice? First, we need to understand whether having more questions pertaining to the concept of integrity has an effect on the results. Second, we need to understand whether we would still receive more empirical support for integrity than availability or confidentiality if a similar study was conducted with a larger sample of participants. This would enable us to know if the study participants' views on security generalize to other digital repository staff members. Third, we need to understand what impact digital repository staff members' views on security actually have on the security of digital repositories. For example, if the principle of integrity is more salient in digital repository staff members' minds, does this mean that digital repositories are less secure when it comes to availability and confidentiality? In other words, are digital repository staff members focusing on integrity at the expense of availability or confidentiality? This may not be the case. It could simply be that integrity is more important than availability or confidentiality. Or it could be that performing actions related to integrity indirectly address issues relating to availability and confidentiality. Or it could be that digital repository managers find it easier to address availability and confidentiality relative to integrity, and so they focus on integrity. At any rate, future research should seek to address these issues so that we can have a better understanding of how what digital repository staff members think about security affects the security of digital repositories.

This study makes two primary contributions to the digital preservation literature. First, it complements the development of standards for TDRs by focusing on the security criteria within one of those standards – DIN 31644. This study examines these security criteria from digital repository staff members' points of view. Prior to this study, we only had the security criteria without insight into the perspectives of those who are responsible for actually addressing those criteria. Second, this study also contributes to the digital preservation literature by providing both qualitative and quantitative data collection instruments which can be used to understand digital repository staff members' perceptions on security. Since efforts to certify trustworthy digital repositories are well underway, and security is a critical element of becoming certified, we anticipate that better understanding digital repository staff members' perspectives on security will only increase in importance going forward.

This study also makes one main contribution to the computer science literature pertaining to security. It takes a classic definition of security, one underpinned by the principles of confidentiality, integrity, and availability, and moves that definition forward by operationalizing the concept with measurement items in a survey instrument. This instrument, what we call the Security Perception Survey (SPS), represents a security metric focused on the perceptions of those responsible for managing and securing computing infrastructures. While SPS was developed using the responses of people who manage and secure TDRs, one specific type of computing infrastructure, subsequent studies could assess the generalizability of SPS to provide insights into the perceptions of people who are responsible for managing and securing other types of computing infrastructures.

The primary limitation of this study is its sample size. Only two digital repository staff members participated in this study. Thus, we cannot generalize the results of this study beyond our sample. However, we felt that who participated in this study was more important than how many. We needed individuals who were at organizations where third parties had verified the success of their security efforts. We felt these individuals would provide the most insightful information about their views on security. We also thought that staff at organizations that successfully passed repository certification by the DIN 31644 standard would be in the best position to evaluate the security criteria within the standard. These issues guided our choices regarding who was eligible to participate in our study, which in turn, led to a small sample size. Despite our small sample size, we reached 100% of our sampling frame; representatives from all of the organizations that have acquired nestor seals of approval participated in this study. It is also important to note the challenges to employing traditional research methods, such as interviews and surveys, to study security. For example, people are reluctant to participate in security studies because: 1) they have concerns about whether the information they provide could somehow be used by others to compromise their systems, or 2) they fear their own shortcomings with respect to their expertise might become exposed as a result of participation [18]. Although we faced a number of these well-documented challenges to recruiting participants for our study, we were yet able to successfully recruit individuals from both organizations that recently acquired nestor seals of approval.

6. CONCLUSION

Security is a major issue for digital repositories. Digital repository staff members are responsible for managing and securing digital repositories, thus their perspectives on security are critically important to understand. This study provided a preliminary investigation into digital repository staff members' views on security and security criteria in standards for TDRs, in particular DIN 31644 and the nestor explanatory notes for Trustworthy Digital Archives. Participants articulated their views on security in terms of integrity and to a lesser extent availability and confidentiality. Results of this study warrant a closer correspondence between research on security in digital preservation and computer science, because of the overlap that results of this study have demonstrated. Participants in this study found the security criteria in the standard that they chose sufficient. Going forward, researchers should continue analyzing digital repository staff members' views on security and security criteria, so that the digital preservation community can validate the relevance and importance of the security criteria by those who are responsible for making digital repositories secure.

7. ACKNOWLEDGMENTS

We thank Michael Frisby and his colleagues at the Indiana Statistical Consulting Center for providing assistance with data analysis and reading prior drafts of this work.

8. APPENDICES

8.1 Appendix A – Interview Protocol

1. How do you define repository trustworthiness? In other words, what does it mean to you for a repository to be trustworthy?
2. How do you define security as it relates to digital repositories? In other words, what does security mean to you?
3. How would you describe the relationship between the trustworthiness of a digital repository and the security of that digital repository? In other words, how would you describe the relationship between security and trustworthiness?
4. Take a minute to read over C34, the nestor criterion on security. Now think back to when you were preparing for audit. How easy or difficult was it to address criterion C34 for your digital repository?
5. How much time do you think it took you and your colleagues to address criterion C34?
6. How prepared were you and your colleagues to address criterion C34?
7. Do you think your repository is more secure as a result of addressing criterion C34? Why or why not?
8. Do you think criterion C34 sets the bar too high for addressing security issues? Or do you think criterion C34 sets the bar too low for addressing security issues? Or do you think criterion C34 sets the bar "just right" for addressing security issues? Why or why not?
9. Do you think any additional criteria should be added to criterion C34 to make digital repositories more secure and therefore more trustworthy? If so, how would you describe what criteria should be added?
10. Did you use DRAMBORA to help you address the security criteria in DIN 31644? If so, which parts of DRAMBORA were most helpful and why?
11. Is there anything else you'd like to add, given our topic of security of Trustworthy Digital Repositories?

8.2 Appendix B – Security Perceptions Survey

Questions pertaining to confidentiality (Questions were answered on a 5-point, likert-type scale ranging from "Strongly disagree" to "Strongly agree" with one additional option: "Not applicable.")

1. Access control mechanisms should be used to support confidentiality (e.g., cryptography).
2. Mechanisms should be used to prevent illicit access to information.
3. The existence of data should be denied to protect it.
4. Resources should be hidden to protect them.

Questions pertaining to integrity (Questions were answered on a 5-point, likert-type scale ranging from "Strongly disagree" to "Strongly agree" with one additional option: "Not applicable.")

1. Improper changes to data should be prevented.
2. Unauthorized changes to data should be prevented.
3. Information about the source of data should be protected.
4. Unauthorized changes to information about the source of data should be prevented.
5. Prevention mechanisms should be used to maintain the integrity of data by blocking any unauthorized attempts to change the data.
6. Prevention mechanisms should be used to maintain the integrity of data by blocking any attempts to change the data in unauthorized ways.
7. Detection mechanisms should be used to report when the data's integrity is no longer trustworthy.
8. System events (e.g., user or system actions) should be analyzed to detect problems.
9. The data itself should be analyzed to see if it has been changed.
10. A system should report what causes integrity violations.
11. A system should report when a file is corrupt.

Questions pertaining to availability (Questions were answered on a 5-point, likert-type scale ranging from "Strongly disagree" to "Strongly agree" with one additional option: "Not applicable.")

1. A system should guard against denial of data attacks.
2. A system should guard against denial of service attacks.
3. An unavailable system is at least as bad as no system at all.

4. A system administrator should be able to tell the difference between when data is not available due to circumstances in the environment versus a security attack.

9. REFERENCES

[1] Bishop, M., 2003. *Computer security : art and science*. Addison-Wesley, Boston.

[2] Butler, S.A. and Fischbeck, P., 2002. Multi-attribute risk assessment. In *Symposium on Requirements Engineering for Information Security*. http://openstorage.gunadarma.ac.id/research/files/Forensics/OpenSource-Forensic/MultiAttributeRiskAssesment.pdf

[3] Cherdantseva, Y. and Hilton, J., 2013. A reference model of information assurance & security. In *Availability, reliability and security (ares), 2013 eighth international conference on* IEEE, 546-555. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6657288&isnumber=6657192

[4] Deutsches Institut Für Normung, 2012. *Information and documentation—criteria for trustworthy digital archives*. Deutsches Institut für Normung. http://www.din.de/en/getting-involved/standards-committees/nid/standards/wdc-beuth:din21:147058907

[5] Digital Curation Centre and Digital Preservation Europe, 2007. *DCC and DPE Digital Repository Audit Method Based on Risk Assessment, v1.0*. http://www.repositoryaudit.eu

[6] Dillo, I. and De Leeuw, L., 2014. *Data Seal of Approval: Certification for sustainable and trusted data repositories*. Data Archiving and Networked Services. http://www.datasealofapproval.org/media/filer_public/2014/10/03/20141003_dsa_overview_defweb.pdf

[7] Herrmann, D.S., 2007. *Complete guide to security and privacy metrics : measuring regulatory compliance, operational resilience, and ROI*. Auerbach Publications, Boca Raton.

[8] Howard, M. and Leblanc, D.E., 2002. *Writing Secure Code*. Microsoft Press.

[9] International Organization for Standardization, 2012. Space data and information transfer systems: audit and certification of trustworthy digital repositories International Organization for Standardization. http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510

[10] Jansen, W., 2009. *Directions in Security Metrics Research*. National Institute of Standards and Technology. http://csrc.nist.gov/publications/nistir/ir7564/nistir-7564_metrics-research.pdf

[11] Nestor Certification Working Group, 2013. *Explanatory notes on the nestor Seal for Trustworthy Digital Archives*. http://files.dnb.de/nestor/materialien/nestor_mat_17_eng.pdf

[12] Organization for Economic Co-Operation and Development, 2002. *OECD Guidelines for the Security of Information Systems and Networks: Towards a Culture of Security*. Organization for Economic Co-operation and Development. http://www.oecd.org/sti/ieconomy/15582260.pdf

[13] Parker, D.B., 1998. Fighting computer crime : a new framework for protecting information J. Wiley, New York :.

[14] Pfleeger, S. and Cunningham, R., 2010. Why measuring security is hard. *IEEE Security & Privacy*, 4, 46-54. http://doi.ieeecomputersociety.org/10.1109/MSP.2010.60

[15] Pfleeger, S.L., 2000. Risky business: what we have yet to learn about risk management. *Journal of Systems and Software 53*, 3, 265-273. doi:10.1016/S0164-1212(00)00017-0

[16] Sen, S. and Samanta, S., 2014. Information security. *International Journal of Innovative Research in Technology*, 1(11), 224-231.

[17] Stoneburner, G., Hayden, C., and Feringa, A., 2004. *Engineering Principles for Information Technology Security (A Baseline for Achieving Security), Revision A*. National Institute of Standards and Technology. http://csrc.nist.gov/publications/nistpubs/800-27A/SP800-27-RevA.pdf

[18] Sundaramurthy, S.C., Mchugh, J., Ou, X.S., Rajagopalan, S.R., and Wesch, M., 2014. An anthropological approach to studying CSIRTs. *IEEE Security & Privacy*, 5, 52-60. doi:10.1109/MSP.2014.84

Will Today’s Data Be Here Tomorrow? Measuring The Stewardship Gap

Jeremy York
University of Colorado Boulder
510 Miller Ave
Ann Arbor, MI 48103
jeremy.york@colorado.edu

Myron Gutmann
University of Colorado Boulder
483 UCB
Boulder, CO 80309-0483
myron.gutmann@colorado.edu

Francine Berman
Rensselaer Polytechnic Institute
110 8th Street
Troy, NY 12180
bermaf@rpi.edu

ABSTRACT

Stakeholders in scholarly research are paying increased attention to stewardship of digital research data[1] for the purposes of advancing scientific discovery, driving innovation, and promoting trust in science and scholarship. However, little is known about the total amounts, characteristics, and sustainability of data that could be used for these purposes. The Stewardship Gap is an 18-month project funded by the Alfred P. Sloan Foundation to understand issues in defining metrics for and measuring the stewardship gap: the potential gap between the amount of valuable data produced through sponsored projects in the United States and the amount that is effectively stewarded and made accessible. This paper reports on the first phase of the project, which sought to develop an instrument to gather information about research data sustainability from a broad variety of researchers and research disciplines and make progress toward the ultimate goals of 1) shedding light on the size, characteristics, and sustainability of valuable sponsored research data and creative work in the United States, and 2) recommending actions stakeholders can take to address the stewardship gap if one is found to exist.

Keywords

Digital curation, digital preservation, research data, data stewardship, data sustainability

1. INTRODUCTION

The explosion of digital information and the promise of using (and reusing) data to spur research innovation have focused attention in the past couple of decades on issues of appropriately curating, managing, and disseminating digital data for reuse. This is true in both the public and private sectors, where digital data are increasingly seen as an asset to be used to promote innovation, economic growth and trust in or accountability of government [12, 13, 40, 48, 58], and to further the arts and advance and verify scientific discovery [5, 14, 18, 36, 37, 38, 40, 55, 62].

Despite high interest in reuse of research data in the scholarly community, numerous challenges have inhibited the ability to understand the size and breadth of the research data universe, or to develop means to ensure that all data “of value” will be discoverable and usable at the appropriate time in the future. Challenges range from difficulty in defining the data of interest [49] and difficulty making measurements (e.g., due to the time required, complexity of social and technical factors, or lack of methods) [1, 8, 16, 17] to challenges in comparing results of different studies [4, 8, 15], poor understanding of the interplay between the many factors involved in data stewardship, and lack of knowledge about how to interpret what has been measured [1, 11].

Concerns that valuable data may not be adequately preserved come in part from studies such as “Sizing the Problem of Improving Discovery and Access to NIH-Funded Data,” [49] which found that 88% of articles published in PubMedCentral in 2011 had “invisible datasets” (where deposit of data in a recognized repository was not explicitly mentioned). Other surveys and studies in recent years have similarly discovered small percentages of data deposited in public repositories. These studies have also uncovered information about data management and storage practices that raise concerns about data persistence, such as lack of future planning for preservation of project data, and significant amounts of research data archived on personal devices as opposed to institutional or community infrastructure [See for example 3, 6, 22, 29, 30, 39, 41, 44, 50, 56, 61].

The Stewardship Gap project was undertaken to investigate means of gathering information about these concerns. In particular, it aims to better understand the potential gap between the total amount of valuable data resulting from sponsored projects in the US that is being produced and the amount that is or will be responsibly stewarded and made accessible to others.

2. LITERATURE REVIEW

2.1 Stewardship Gap Areas

We conducted a survey of relevant literature to ascertain what is known about the stewardship gap. This survey revealed the presence of not one, but many gap areas that impact the ability to measure, analyze, plan for, and act to steward valuable research data. These areas include:

1. Culture (e.g., differences in attitudes, norms and values that affect data stewardship and reuse);
2. Knowledge (e.g., about how to preserve data, what skills are needed, how much data exist, of what kind, how much of it has value and for how long);
3. Commitment (e.g., commitments adequate to needs for stewardship and reuse);
4. Responsibility (e.g., who is responsible for funding stewardship and carrying out stewardship activities);
5. Resources (e.g., funding, infrastructure, tools, human resources);
6. Stewardship actions such as curating, managing, and preserving data, and activities that enable curation, management, and preservation such as making data available (e.g., through data sharing or deposit in a data repository), long-term planning, and collaboration [64].

While all of these areas appeared crucial to understanding the stewardship gap as a whole, we designed a pilot project that would provide evidence of the presence of a gap and important elements of any gap we discovered. Based on background reading and focused interactions with the project advisory board, we hypothesized these elements to be the extent and duration of value that data have and the extent and duration of commitments made to steward valued data. We considered that if our study found e.g., that a given dataset had value for twenty years, but there was only a commitment to preserve the data for five, this could be an indication of a stewardship gap. We believed information about value and commitment would have greater value if combined with information about who could act to address a gap if one existed, and thus added stewardship responsibility as a third primary parameter in the study.

The first phase of the study was devoted to formulating questions about research data value, commitment, and stewardship responsibility that could be answered by researchers in a wide variety of disciplines about data of diverse sizes and types. Research in the first phase focused on data resulting from public- or non-profit-sponsored research conducted at institutions of higher education in the United States.

2.2 Review of Research Data Studies

There are two main types of studies that have sought to measure aspects of the stewardship gap for research data. The first comprises studies with a specific focus (“targeted” studies), for instance on research data sharing, deposit of data in repositories, or funding for stewardship [some examples include 23, 43, 46, 47, 54, 63]. Studies of the second type (“wider” studies) cover a range of topics at once, often at less depth for any given topic than a targeted study [e.g., 3, 21, 24, 26, 34, 35, 39, 41, 44, 45, 56, 59]. Many of the second type were conducted on university campuses to gather information to help establish or improve research data management services, though some studies extended across campuses as well [e.g., 18, 30].

Figure 1 shows the distribution of one hundred-seven studies reviewed for the stewardship gap project in the six gap areas described above. Studies related to data value are included under “Culture.” However, they are also represented in the figure as a separate category since value is a main focus of the project. The figure also shows the number of targeted versus wider studies.

Figure 1 is not a comprehensive representation of all studies related to the stewardship gap. It does show the topical distribution among a significant subset, however, and shows in particular how our three areas of interest (data value, stewardship commitment, and stewardship responsibility) are represented in the broader landscape of studies.[2]

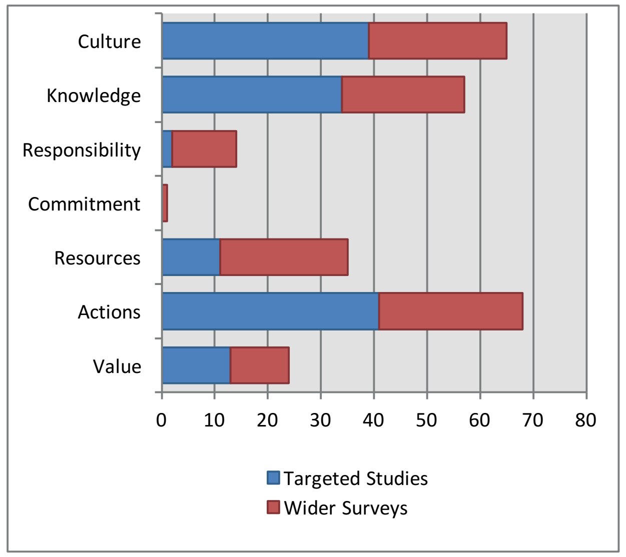


Figure 1. Prior measures of stewardship gap areas

2.3 Value, Commitment, and Responsibility

2.3.1 Value

We identified two general types of studies related to data value. The first focuses on means of understanding the value and impact of data and data stewardship, often in financial or business terms [some examples are 7, 9, 10, 20, 25, 28, 29, 31, 32, 51, 52, 57, 60]. The second type, which is most relevant to the stewardship gap project, investigates different kinds, degrees, or durations of data value. One example is Akmon’s 2014 investigation of how scientists understand the value of their data throughout the course of a project and the effect of conceptions of value on data practices [2]. Two other, wider ranging studies of this type[3] include a campus study that asked researchers whether their data would be valuable for projects other than the one they were gathered for (though the study did not ask about users of data or or reasons for data value) [50] and the PARSE.Insight project [30], which asked respondents to rate the importance of the following “well-known” reasons for data preservation:

1. If research is publicly funded, the results should become public property and therefore properly preserved
2. It will stimulate the advancement of science (new research can build on existing knowledge)
3. It may serve validation purposes in the future
4. It allows for re-analysis of existing data
5. It may stimulate interdisciplinary collaborations
6. It potentially has economic value
7. It is unique

The Stewardship Gap’s investigation of value is most similar to the PARSE.Insight project in that it poses a range of different reasons for data value for researchers to respond to. It differs, however, in asking researchers to rate reasons for value, as value. However, early pilot studies encountered difficulty classifying value according to criticality of data to the institution as the framework specified [26, 34]. Explicit questions about data value do not appear to have been included in subsequent implementations, although there are questions about whether data should be preserved, whether they can be reused [3, 21], and how long data will be archived [3, 39, 45, 61]. Because they gather data that provide indicators of data value, these studies have been included in the tally of studies shown in Figure 1.

[1] Unless otherwise indicated, “data” and “research data” are used to refer to digital data throughout the paper, as opposed to analog data.

[2] See https://www.zotero.org/groups/data_stewardship_studies/items for a list of all studies represented.

[3] Several additional studies have been undertaken that used or were based on the Digital Asset Framework, a framework developed by the Humanities Advanced Technology and Information Institute at the University of Glasgow in association with the Digital Curation Centre to conduct an assessment of data assets in an institutional context. In its early instantiation, the DAF framework was designed to gather information about data

opposed to reasons for preservation—though our study does also ask researchers to indicate reasons for value that have had the greatest impact on decisions about preservation.

2.3.2 Commitment

Commitment is significantly understudied in comparison with other areas. Only one study was found that gathered information about commitment.[4] This was a 2005 survey conducted by Tessella for the Digital Preservation Coalition as part of an initiative to assess digital preservation needs in the United Kingdom [61]. The survey asked whether there was a high level of commitment to digital preservation in the respondent's organization. In contrast to this study, the Stewardship Gap asks about levels of commitment associated with research data from specific projects.

2.3.3 Responsibility

A number of the reviewed studies investigated questions of responsibility, including responsibility for:

- Storage, backup, and management of research data
- Funding of data management, storage, and preservation
- Decisions about stewardship, including which data are important to preserve and for how long, what constitutes compliance with regulations, licenses, and mandates, what descriptive metadata are appropriate, and provisions for short- and long- term storage and preservation [21, 26, 30, 33, 39, 45, 50, 61]

The Stewardship Gap did not introduce new questions in this regard. However, our primary purpose was to be able to compare information about responsibility with information about value and commitment in order to understand who could act to address a stewardship gap if one existed.

2.4 Common Themes In Reviewed Literature

Some common themes across reviewed studies that were relevant to our efforts to develop a strategy for measuring the stewardship gap included the following:

1) The significant amount of time that can be involved in conducting a study. Many studies employed a preliminary pilot phase to refine questions, followed by a broader survey and then follow-up interviews. This was done to balance the needs to survey a sufficiently large population but also gain important contextual information that interviews can provide.

2) The challenges of creating a common understanding of what "data" are for the purposes of the study. Three main challenges that surfaced were 1) addressing what type of materials are included in "data" (e.g., course materials, notes, structured data, images, non-digital files, etc.), 2) what constitutes a "dataset" (e.g., an individual file, a set of files in a particular format, or a set of related data regardless of format), and 3) what universe of data is being measured (e.g., all data held or created, all data held or created in a specific time frame, or data from a specific project).

3) The significance of the correlation between research discipline, research role, type of research (e.g., scientific or clinical), and level of experience and the amount of research data generated, attitudes and practices about data storage, data sharing and reuse, and beliefs about the primary reasons for and threats to preservation.

4) The broad diversity in sizes and formats of digital data generated and types of storage used, and the relatively small amounts of data that are deposited with disciplinary or other repositories outside the researcher's institution [30, 39, 41, 59].[5]

Regarding this last theme, the University of Nottingham investigated their results further and found that the majority of researchers stored their data in multiple locations [41]. This would seem to add a degree of confidence to concerns about adequate preservation of data. However, the Open Exeter Project found that much of the research data being held is not actively managed, raising additional concern [39]. This concern is supported by results from the University of Northampton that while most researchers intend to keep data beyond the completion of a project, and even indefinitely, this intention is not realized for a variety of reasons, including

- lack of data management strategies
- the need to store files that exceed computer hard-drive space on external media that are more prone to degradation and loss
- files and software stored on personal devices becoming out of sync with university resources that are needed to access and use them [3].

The considerations these common themes raised for our project and the ways we addressed them are described in section 3 below.

3. GOALS AND METHODOLOGY OF THE STEWARDSHIP GAP PROJECT

The first goal of our initial study was to test, across as broad a range of disciplines as possible, the performance and effectiveness of questions about data value, stewardship commitment, and stewardship responsibility. Because we wanted to be able to gather information about a measureable "gap", we also wanted to collect information about the size and characteristics of valued data. The second goal was to analyze responses in order to inform a more in-depth study of the stewardship gap in a second phase.

To accomplish these goals, we designed a questionnaire (see the question areas in Table 1) and conducted interviews with seventeen researchers in sixteen fields from thirteen US institutions over the course of November and December 2015. Interviewees were selected on the basis of their association with at least one of a range of academic disciplines, with the goal of achieving a wide range of disciplinary coverage. Most of the interviewees were known to or suggested by members of the project team or advisory board. Overall, thirty-one researchers were contacted, yielding a response rate of 55%.

3.1 Methodological Considerations

Some important considerations and decisions have made our study both similar to and different from preceding studies. These include:

1) Our study was preliminary, and in the context of other studies would fall into the preliminary pilot stage. The questions we developed were drawn out of our literature review and initial discussions with the project advisory board. We centered the questions around issues of value, commitment, and responsibility, and then added questions relevant to other gap areas (e.g., infrastructure, sustainability planning) as they supplemented and supported these focal areas. Gathering relevant information in the least amount of time was a primary goal.

2) We decided to target project principle investigators (PIs) as subjects. We realized that PIs might describe their data and the way it is managed differently than others involved in the project,[6] but were concerned with learning about data value, stewardship commitments, and responsibility for stewardship and believed PIs to be primary sources for this information.

Table 1. Stewardship Gap Question Areas

Research Context	What is the purpose of the project? What domains of science or creativity are the resulting data in? Who are the project collaborators and funders? What are the characteristics and what is the overall size of the project data?
Commitment	For how much of the data is there: a formalized commitment to preserve; an intention to preserve; no intention to preserve (though no intention to delete); the data are temporary (and will be deleted)?
Stewardship	Who is currently stewarding the data? What is being done to take care of the data? Are there any concerns about the ability to fulfill the intention or commitment? What prospects exist when the current commitment or intention is over?
Value	Why are the data valuable and for how long? How does the valuation affect stewardship decisions? Would it be worthwhile to reassess the value of the data in the future?

3) We asked PIs about data from a single project, rather than data from all projects that they might be responsible for. This decision was made in order to have a coherent view of what it is we were discussing with researchers: a single research project, however broadly that might be defined. We asked researchers in particular to describe data from a project of their choosing where the project was:

- Funded by a public- or non-profit source
- One for which they were responsible for generating the digital data or creative content
- One for which they were able to speak confidently about questions of size, content characteristics, and preservation commitments related to the data.

4) We did not present interviewees with any definitions or parameters for understanding "data". As a pilot study, our concern in this and other areas was to hear researchers answer from their own perspective about the questions we raised (although we did define Steward and Preserve, two terms that were important to our framework for measuring commitments on data).[7]

5) For the purposes of analyzing results, we treated "datasets" as the researcher defined them. For instance, if a researcher designated three different datasets, one each of interviews, field samples, and GIS information, we understood these to be three datasets, regardless of the formats or types of data included in each.

6) We asked about specific types of value and value duration, and researchers' agreement with whether specific types applied to their data. We presented categories of value, but also gave researchers the opportunity to add their own (see Table 3 below, and following).

7) We asked researchers to place the data generated in their projects into one of four categories of commitment, associating a term of commitment with each where applicable. We choose these categories specifically to distinguish between formal and informal commitments on research data, and to look for patterns in the association of specific types of value with types and durations of commitment. The four categories are given in Table 1.

4. RESULTS

4.1 Research Context

Seventeen PIs were interviewed for the study. In the seventeen projects they described, PIs provided information about value and stewardship commitments on a total of 40 datasets. Table 2 shows the distribution of researcher fields, the number of datasets described in each area, total size of the datasets, and whether datasets included sensitive information (information that is private, proprietary, or confidential). Excepting environmental studies where two researchers were interviewed, only one researcher was interviewed in each discipline.

Table 2. Research Discipline and Dataset Details

Researcher Discipline	Number of Datasets	Size of all datasets	Sensitive data
Geography	5	< 5 GB	None
History	6	< 5 GB	None
Archaeology	2	< 5 GB	--[8]
Economics	1	< 5 GB	All
Political science	2	< 500 GB	A portion
Psychology	1	< 20 TB	A portion
Public administration	3	< 100 GB	All
Information	3	< 500 GB	A portion
Education	2	< .1 GB	A portion
Environmental studies	6	< 500 GB	A portion
Human physical performance and recreation	1	< 100 GB	A portion
Neuroscience	2	< .1 GB	None
Astronomy	1	< 50 TB	For a time[9]
Computer sciences	1	< .1 GB	None
Physics	3	< 50 TB	A portion
Statistics	1	< 500 GB	None

[4] A second study included metrics related to stewardship commitment [42], but did not undertake measurement.

[5] The PARSE.Insight project found that 20% of respondents submitted data to a data archive [30]; a University of North Carolina study found this number to be 17% [59]; at the University of Nottingham 2% of respondents said they stored data in an institutional repository (the only repository option) [41]; at the University of Exeter about 4% indicted they deposited data in a public repository when they have finished with it [39].

[6] Two studies [45, 46] found differences in data descriptions by principle investigators and researchers, and a third [31] found data created by researchers that were not passed on to data managers.

[7] We defined Steward as "to responsibly manage data that is in your care (including the wide variety of activities that might be involved in managing them)" and Preserve as "to execute a set of activities with the explicit goal of maintaining the integrity of data over time."

[8] Did not ask

[9] Data were restricted during a time of analysis, and then released to the public.

Five projects reported no sensitive information in resulting data, eight included some sensitive information or were restricted for a time, and all data were sensitive in two projects. In only one project where a portion of data were sensitive did a researcher make an explicit distinction between the value associated with the sensitive data and the non-sensitive data.

The start and end dates of the investigated projects spanned from 1948 to 2018 with most projects taking place between 2000 and 2015. Some of the projects had been continuously funded for decades, some were completed, and some were still ongoing. Many of the projects were conducted in multiple phases with funding from different sources, and some were continuations of or components of other projects. Despite these complexities, researchers did not have trouble identifying the specific data associated with the projects they selected for the interviews. Regardless of time, the number of funders, or changing collaborators, researchers had a strong sense of a cohesive activity that they viewed as a project, and its associated data assets.

We wished to cause as little disruption to researchers as possible and therefore notified them in advance that no research into the details of their data were required prior to the interview. We asked about details nonetheless in order to gauge what might be required to obtain this information if desired in a more in-depth study. We found that difficulty describing sizes and attributes of data varied across respondents. Many knew approximate sizes and formats offhand or had the information readily available during the interview. Others had a strong sense of what data were collected or produced (e.g., interview transcripts, images, etc.) but could not recall specific details.

A related issue we encountered, experienced in previous studies as well, resulted from researchers' understandings of what were considered "data". In most interviews, the researcher's description of his or her data evolved over time, as they remembered additional sets of data or provided more information in order to answer subsequent questions (for instance about the stewardship environment). In a few cases, however, we found that certain sets of data were not described initially because they were not considered as "project data" by the researcher. Some of these types of data included images taken on-site during field studies, audio of interviews, data that are produced as primary data are analyzed and refined, descriptive and contextual information about the data, and video recordings of study participants.

Whatever their challenges in remembering the details about data, interviewees had little difficulty answering questions about commitments on data, data value, or responsibility for stewardship.

4.2 Type of Value

The interview asked researchers to indicate their degree of agreement with eighteen different types of value (see Table 3). The degree choices were strongly agree, agree, neutral, disagree, and strongly disagree. Researchers could also indicate they were unsure or that the type of value did not apply, or specify "other" types of value. The eighteen value types can be grouped into four main categories, as shown in Table 3: value due to 1) reuse potential, 2) cost of reproduction, 3) impact, and 4) scholarly practice.[10] We also asked questions about the value of data over time. There were four datasets for which no information about value was obtained and one dataset for which only partial information about value was obtained, primarily due to time constraints on the interviews.

Table 3. Types of Value

Reuse potential: Audience	Value for the researcher's own research
	Value within the researcher's immediate community of research
	Value outside the researcher's immediate community of research
	Broadly applicable value (e.g., as cultural heritage or inclusion in a reference collection)
Reuse potential: Reasons for or factors that affect reuse	Value increases in combination with other data
	Data only has value when combined with other data
	Value due to the organization or usability of the data
	Value due to the timeliness or timely relevance of the data
	Value for use in support services (such as calibrations or search services)
	Value for audit purposes or because the data have been mandated to be kept
Cost	Value because the data would be costly to reproduce
Impact	Value due to demonstrated or potential impact (in terms of people, money, time, policy, transformative potential, or some other factor)
Scholarly practice	The data are retained in conformance with good scholarly practice
Change in value over time	The data have gained value over time
	The data will gain value over time
	The data have lost value over time
	The data will lose value over time[11]
	The data are timeless (will never lose their value)

Some of the "other" types of value respondents mentioned were:

- Historic value
- Value to facilitate research (training data)
- Value to facilitate policy-making
- Use for quotes in outreach
- Use as examples in teaching and executive development
- Repeatability, reference, transparency
- Longitudinal value
- Model for other studies
- Type of study: Resolution and context (moving between individual and societal analysis)

4.3 Type of Value and Term of Value

Figure 3 shows the main categories of value that researchers strongly agreed applied to their data, and the durations over which they believed the data would have value. Value for researchers' own use is represented separately from value for others' use. Reasons for or factors that affect reuse and information about change in value over time are excluded from the chart to focus the results on the high-level value categories investigated.

As Figure 3 indicates, researchers believed much of their data would have value for a long time. They most frequently expressed strong agreement with the value data held for their own research, followed by value due to the cost involved in producing data, value as evidenced by reuse by others (including both in and outside their immediate community), and the demonstrated or potential impact the data could have. Researchers also strongly agreed that they retained their data in conformance with good scholarly practice.

[10] Keeping data for reasons of good scholarly practice is not strictly a type of value. However, there was such strong agreement with this as a reason for keeping data that it is included alongside other results.

[11] Questions about lost value were only asked if respondents were neutral or negative about increase in value.

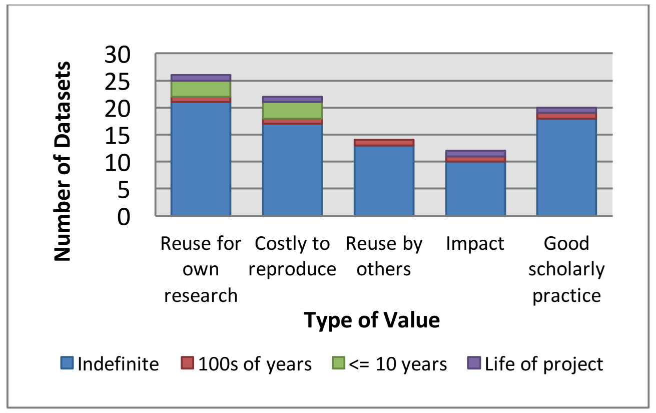


Figure 3. Type of value and term of value

By contrast, when researchers were asked which reasons for value had the greatest influence on decisions about preserving data, demand for data was most frequently cited, with difficulty of reproduction and use for their own research mentioned least frequently. These results show a difference between the types of value researchers most strongly agree that their data have and the reasons for value that have the greatest impact on decisions about data preservation.

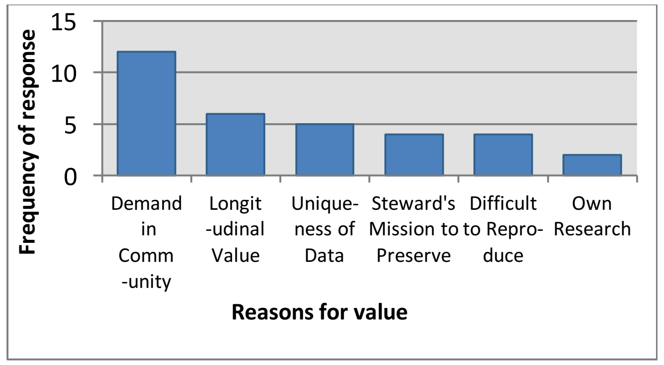


Figure 4. Reasons for value with the greatest impact on decisions about research data preservation

4.4 Commitment and Value

Figure 5 shows the types and terms of commitments that researchers associated with their project data. Our results indicate that researchers have strong intentions to preserve much of their data. While nearly ¾ of datasets carried either an intention or commitment of preservation, however, only two of the twenty datasets desired to be kept more than 10 years had a matching duration of commitment (see Figure 5 – there is a commitment term of more than five years for only two of the five datasets where commitments were expressed).

Juxtaposing type of commitment with term of value (see Figure 6) reveals a similar story, with only 5 out of 37 datasets believed to have value for more than 10 years carrying a commitment of any duration (three of the five commitments are for less than 5 years).

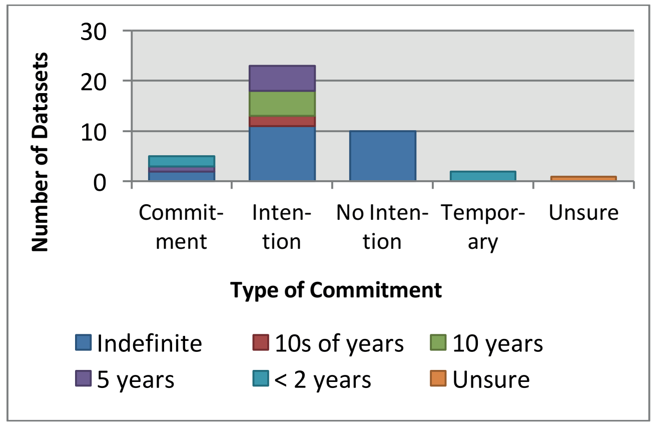


Figure 5. Type of commitment and term of commitment

These results raise an important question about whether intentions to preserve data translate ultimately into preserved data. It is notable that only one quarter of the datasets were identified as having indefinite value, but carried no preservation intention or commitment.

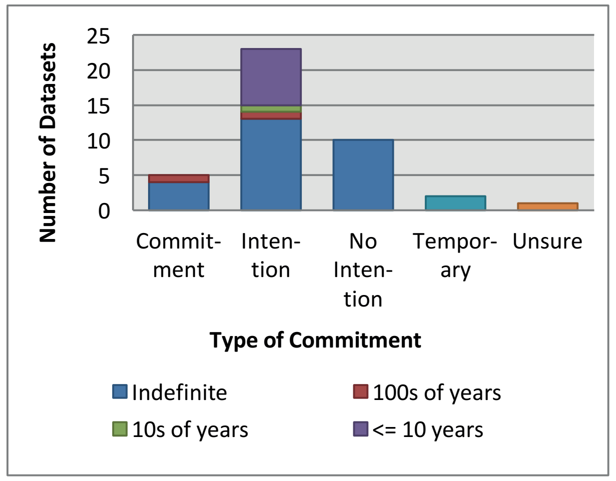


Figure 6. Type of commitment and term of value

4.5 Responsibility

The initial study gathered information about those responsible for funding the creation of research data, and those who have or might have ongoing responsibility for stewardship of the data, whether the role is as a funder or executor of stewardship activities. A tabulation of the most common funding sources for the projects investigated is given in Figure 7. Many of the projects had more than one funder.

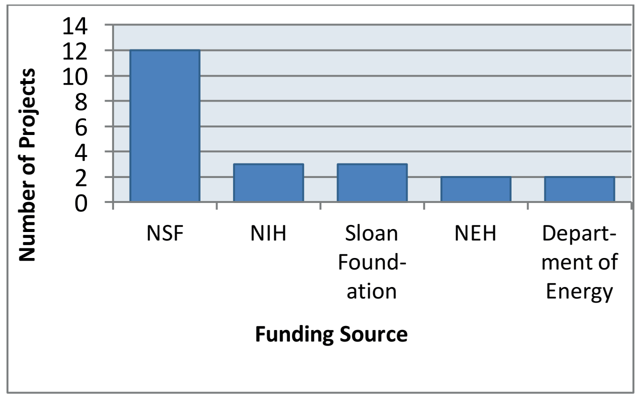


Figure 7. Sources of project funding

In only two of the seventeen projects had project data been transferred to someone other than the one who was originally responsible for the data during the project. Figure 8 shows who researchers indicated was responsible for stewardship, separated into categories of personal stewardship (the data are on a personal computer, removable media, etc.), stewardship within an institution (within a lab or institutional repository) and multi-institutional or public stewardship (e.g., a repository that is operated on behalf of or for use by multiple institutions or the public). The figure also shows responses of researchers when asked how confident they felt in the ability of the person or entity stewarding the data to fulfill the commitment on intention that existed on the data.

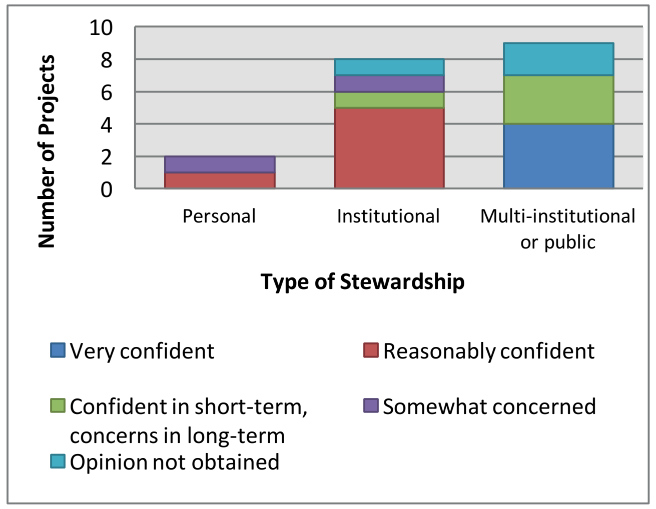


Figure 8. Responsibility and confidence in stewardship

No responses were obtained for three of the projects, but responses from the remaining projects were somewhat mixed. There is high confidence in multi-institutional or public repositories, but also concern about funding beyond the near term. There is reasonable confidence in institutional solutions, but also concern, including about long-term funding. There is both confidence and concern related to personal stewardship.

Some important questions in interpreting these results are the degree of knowledge researchers have about the environments where their data are stewarded, and how well founded their confidence is. No trend emerged in our interviews regarding the former. Some researchers displayed exceptional knowledge about the stewardship environments for their data while others were less knowledgeable, both about environments and which departments or staff were managing the data.

The question of confidence is also complicated, and relates to the issue of intentions translating into preserved data. As noted earlier, there are a number of considerations in determining the adequacy of management and preservation solutions. The ways we have determined to address issues of confidence in the second phase of research are given in the final section of the paper.

In addition to concerns about stewardship during the current period of commitment or intention, our results indicate that attention should be paid to stewardship after the period of commitment or intention is over. While it is a not a new notion that stewardship needs exist after the period of active data use, our results uncover not to a workflow issue (what happens with data when the project is over) but a commitment issue (what happens when the commitment or intention on the data is over). Responses to the question of what plans exist when the current commitment or intention is over are shown in Figure 9. We did not ask the question consistently across all interviewees. However, with only one researcher indicating definite plans for stewardship, the question bears broader investigation. Even if it could be demonstrated that data were secure while a researcher is active in the field, what plans for valuable data exist after the researcher has retired or no longer has an intention to keep the data?

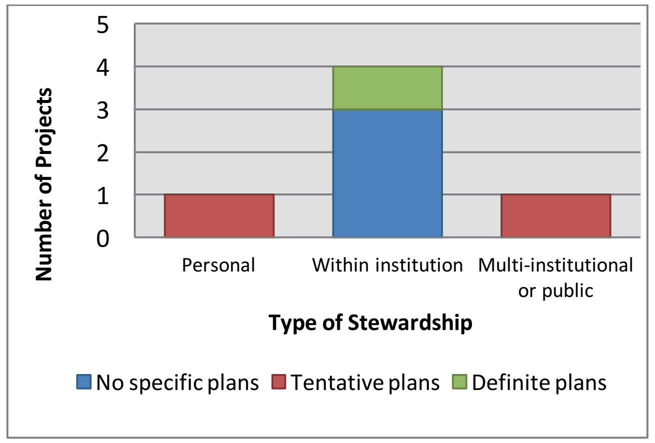


Figure 9. Stewardship plans after existing commitment is over

4.6 Size of Dataset and Data Value

We did not find any correlation between the size of data and the type or term of value, or the size of data and confidence in stewardship, a result that is relevant to future study of the stewardship gap. Researchers across the spectrum of projects believed strongly in the value of their data whether the data were larger or smaller, and types of value were distributed rather evenly across sizes. Similarly, our results did not show that larger sets of data were taken care of better than smaller sets or vice versa. This result has implications for the development of a strategy for measuring the stewardship gap going forward. If it holds true for a larger sample of projects, investigation of the stewardship gap should recognize the value and impact of large data and small alike.

5. DISCUSSION AND FUTURE WORK

5.1 Discussion

The preliminary interviews concluded with questions about whether the interview allowed respondents to describe their data in a way that was meaningful and accurate from their point of view, and what difficulty, if any, they experienced in answering the questions. The response to the first question was unanimously positive in fifteen of seventeen cases where it was asked,[12] and only minor difficulty was reported by two respondents in answering questions about commitment.

These results suggest that we successfully fulfilled the first goal of the initial study, which was to develop an instrument capable of gathering information about data value, stewardship commitments, and responsibility for stewardship across a range of disciplines.

The second goal of the study was to obtain information to inform a more in-depth study of the stewardship gap in a second phase, and our results provided this as well. In particular, they uncovered three main sets of questions, given below.

The first set of questions relates to whether the areas investigated are adequate to provide indicators of a stewardship gap. For example, preliminary results indicate that there could be a large amount of data regarded as being high in value that lack a

[12] We did not receive a response for two interviews due to time constraints.

sufficient commitment for stewardship. If this result were borne out in a larger sample of projects, however, would it point to a meaningful gap (that is, is there cause for concern or do intentions to preserve data in fact result in valuable data being preserved for future use)? What information might be needed to clarify or confirm this?

Our results also indicate a lack of correspondence between particular stewardship environments and confidence in stewardship. What additional information might be needed to validate researcher confidence, or assess the strength of existing stewardship environments?

A second set of questions relates to the selection of an appropriate sampling frame for a more structured study. A lack of correlation between data size on one hand and data value and confidence in stewardship on the other indicate a need to include a diversity of data sizes. How should a sample be structured to do this? Previous studies have found that discipline, researcher role, and level of experience have an impact on many factors such as amounts of data generated, attitudes about data sharing and management and preservation practices. Do all of these variables need to be represented in a sampling frame to provide meaningful results, and if so how can they be best represented?

A third set of questions relates to the granularity of information gathered. For instance, we did not investigate responsibility at the depth of some of the previous surveys (some of which included specific responsibility for data storage, management, etc.). However, the personal, disciplinary, institutional, and multi-institutional dimensions of responsibility appear to be appropriate high-level indicators of who can act to address a stewardship gap if it exists. Is this correct and do these levels provide adequate guidance as to who should act to address a gap if one is found to exist?

Similarly, the level of detail we obtained about data sizes and attributes appears to have been sufficient to associate distinct data with specific commitments, types of value, and responsible entities, our core indicators for determining the presence of a stewardship gap. However, many researchers were not entirely confident in their representation of specific formats used and sizes of data. How accurate do the descriptions of data need to be to provide a meaningful characterization of the stewardship gap?

As other studies have found, there is a direct relationship between the granularity of information that is gathered and the difficulty and amount of time needed to gather it. What is the optimum balance for obtaining meaningful results with minimum imposition on respondents?

5.2 Implications for Future Work

In light of the preceding questions and what we have learned from phase 1 of the study, the following are the major decisions and modifications we intend to make to the instrument in the second phase:

To further address the question of whether intentions translate into commitments for data stewardship we will clarify the purpose of projects, recognizing that there can be a difference between projects with a focus on data creation (e.g., with the explicit purpose of sharing with other researchers) and those where data are not explicitly intended for sharing. We will also gather information about what researchers expect will happen to their data when the current intention or commitment to preserve the data is over, and seek to better understand researchers' goals when transferring responsibility for data to others.

In the development of a sampling frame, we intend to focus first and foremost on stratification by researcher discipline and funding source. This is due to time considerations and the need to prioritize certain variables to keep interviews to a reasonable length. Additional factors may need to be explored more in a further study.

On the question of granularity, we intend to keep to the levels of information we have been gathering about data size and characteristics, stewardship environments (e.g., personal, institutional, multi-institutional), and responsibility for stewardship. While further study may indicate that greater granularity is needed, the current levels appear adequate to our purposes of investigating the presence of a stewardship gap and making general recommendations about how to address it if we find one exists. As above, more detailed analysis may be necessary to make targeted recommendations, depending on the results of the second phase of research.

6. ACKNOWLEDGMENTS

Our deep thanks are due to the Stewardship Gap project advisory board, including George Alter, Christine Borgman, Philip Bourne, Vint Cerf, Sayeed Choudhury, Elizabeth Cohen, Patricia Cruse, Peter Fox, John Gantz, Margaret Hedstrom, Brian Lavoie, Cliff Lynch, Andy Maltz, Guha Ramanathan, and to the Alfred P. Sloan Foundation for their funding of the project.

7. REFERENCES

[1] Addis, M. 2015. *Estimating Research Data Volumes in UK HEI*. http://figshare.com/articles/Estimating_Research_Data_Volumes_in_UK_HEI/1575831

[2] Akmon, D. 2014. *The Role of Conceptions of Value in Data Practices: A Multi-Case Study of Three Small Teams of Ecological Scientists*. University of Michigan.

[3] Alexogiannopoulos, E. et al. 2010. *Research Data Management Project: a DAF investigation of research data management practices at The University of Northampton*. http://nectar.northampton.ac.uk/2736/

[4] Ashley, K. 2012. Generic Data Quality Metrics – what and why. (Arlington, VA, 2012).

[5] Association of Research Libraries Workshop on New Collaborative Relationships 2006. *To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering*. http://www.arl.org/storage/documents/publications/digital-data-report-2006.pdf

[6] Averkamp, S. et al. 2014. *Data Management at the University of Iowa: A University Libraries Report on Campus Research Data Needs*. http://ir.uiowa.edu/lib_pubs/153/

[7] Beagrie, N. et al. 2012. *Economic Impact Evaluation of Research Data Infrastructure*. Economic and Social and Research Council. http://www.esrc.ac.uk/files/research/evaluation-and-impact/economic-impact-evaluation-of-the-economic-and-social-data-service/

[8] Beagrie, N. and Houghton, J. 2014. *The Value and Impact of Data Sharing and Curation*. http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report,_Jan14_v1-04.pdf

[9] Beagrie, N. and Houghton, J. 2013. *The Value and Impact of the Archaeology Data Service: A Study and Methods for Enhancing Sustainability*. Joint Information Systems Committee. http://repository.jisc.ac.uk/5509/1/ADSReport_final.pdf

[10] Beagrie, N. and Houghton, J. 2013. *The Value and Impact of the British Atmospheric Data Centre*. Joint Information Systems Committee. http://repository.jisc.ac.uk/5382/1/BADCReport_Final.pdf

[11] Becker, C. et al. 2011. A Capability Model for Digital

Preservation: Analysing Concerns, Drivers, Constraints, Capabilities and Maturities. (Singapore, Nov. 2011). http://www.academia.edu/1249924/A_Capability_Model_for_Digital_Preservation_Analysing_Concerns_Drivers_Constraints_Capabilities_and_Maturities

[12] Berman, F. et al. 2010. Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. (2010), 110. http://blueribbontaskforce.sdsc.edu/biblio/BRTF_Final_Report.pdf

[13] Big Data Value Association 2015. *European Big Data Value Strategic Research & Innovation Agenda*. Big Data Value Europe. http://www.bdva.eu/sites/default/files/europeanbigdatavaluepartnership_sria__v1_0_final.pdf

[14] Borgman, C.L. 2012. The conundrum of sharing research data. 63, 6 (2012), 1059–1078. http://onlinelibrary.wiley.com/doi/10.1002/asi.22634/epdf

[15] Borgman, C.L. et al. 2014. The Ups and Downs of Knowledge Infrastructures in Science: Implications for Data Management. *Proceedings of the Joint Conference on Digital Libraries, 2014 (DL2014)*. (2014). http://works.bepress.com/borgman/321

[16] Brown, S. et al. 2015. *Directions for Research Data Management in UK Universities*. http://repository.jisc.ac.uk/5951/

[17] Cummings, J. et al. 2008. Beyond Being There: A Blueprint for Advancing the Design, Development, and Evaluation of Virtual Organizations. *Technology*. 3, 2 (2008). http://web.ci.uchicago.edu/events/VirtOrg2008/VO_report.pdf

[18] Fearon, D. et al. 2013. *ARL Spec Kit 334: Research data management services*. Technical Report #9781594079023. Association of Research Libraries. http://publications.arl.org/Research-Data-Management-Services-SPEC-Kit-334/

[19] Federer, L. et al. 2015. Biomedical Data Sharing and Reuse: Attitudes and Practices of Clinical and Scientific Research Staff. *PLoS ONE*. 10, 6 (2015). http://dx.doi.org/10.1371/journal.pone.0129506

[20] Finn, R. et al. 2014. *Legal and ethical barriers and good practice solutions*. Policy RECommendations for Open access to research Data in Europe (RECODE). http://recodeproject.eu/wp-content/uploads/2014/05/D3.1-legal-and-ethical-issues-FINAL.pdf

[21] Gibbs, H. 2009. *Southampton Data Survey: Our Experience and Lessons Learned*. University of Southampton. http://www.disc-uk.org/docs/SouthamptonDAF.pdf

[22] Guindon, A. 2014. Research Data Management at Concordia University: A Survey of Current Practices. *Feliciter*. 60, 2 (2014), 15–17. http://connection.ebscohost.com/c/articles/95923248/research-data-management-concordia-university-survey-current-practices

[23] Hedstrom, M. et al. 2006. Producing Archive-Ready Datasets: Compliance, Incentives, and Motivation.

[24] Hofelich Mohr, A. et al. 2015. Data Management Needs Assessment - Surveys in CLA, AHC, CSE, and CFANS. (2015). http://conservancy.umn.edu/handle/11299/174051

[25] Houghton, J. and Gruen, N. 2014. *Open Research Data Report*. http://ands.org.au/resource/open-research-data.html

[26] Jerrome, N. and Breeze, J. 2009. *Imperial College Data Audit Framework Implementation: Final Report*. Imperial College London. http://ie- repository.jisc.ac.uk/307/

[27] Jones, S. et al. 2008. The Data Audit Framework: a toolkit to identify research assets and improve data management in research led institutions. (London, UK, Sep. 2008). http://www.bl.uk/ipres2008/ipres2008-proceedings.pdf

[28] Kejser, U.B. et al. 2014. *4C Project: Evaluation of Cost Models and Needs & Gaps Analysis*. http://www.4cproject.eu/d3-1

[29] Kroll, S. and Forsman, R. 2010. *A Slice of Research Life: Information Support for Research in the United States*. Technical Report #1-55653-382-9 978-1-55653-382-2. http://www.oclc.org/content/dam/research/publications/library/2010/2010-15.pdf

[30] Kuipers, T. and Hoeven, J. van der 2009. *PARSE.Insight: Insight into Digital Preservation of Research Output in Europe: Survey Report*. http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

[31] Manyika, J. et al. 2011. Big data: The next frontier for innovation, competition, and productivity. (2011), 156. http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

[32] Manyika, J. et al. 2013. Open data: Unlocking innovation and performance with liquid information. (Oct. 2013), 103. http://www.mckinsey.com/business-functions/business-technology/our-insights/open-data-unlocking-innovation-and-performance-with-liquid-information

[33] Marchionini, G. et al. 2012. Curating for Quality: Ensuring Data Quality to Enable New Science. (2012), 119. http://dl.acm.org/citation.cfm?id=2582001

[34] Martinez-Uribe, L. 2009. *Using the Data Audit Framework: An Oxford Case Study*. http://www.disc-uk.org/docs/DAF-Oxford.pdf

[35] Mitcham, J. et al. 2015. *Filling the Digital Preservation Gap. A JISC Research Data Spring Project. Phase One Report*. http://figshare.com/articles/Filling_the_Digital_Preservation_Gap_A_Jisc_Research_Data_Spring_project_Phase_One_report_July_2015/1481170

[36] National Academy of Sciences 2009. Ensuring the integrity, accessibility, and stewardship of research data in the digital age. 325, 5939 (2009), 368. http://www.nap.edu/catalog.php?record_id=12615

[37] National Research Council 2003. *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. National Academies Press. http://www.nap.edu/catalog/10613

[38] National Science Foundation, Cyber Infrastructure Council 2007. Cyberinfrastructure Vision for 21st Century Discovery. *Director*. March (2007). http://www.nsf.gov/pubs/2007/nsf0728/nsf0728.pdf

[39] Open Exeter Project Team 2012. *Summary Findings of the Open Exeter Data Asset Framework Survey*. University of Exeter. https://ore.exeter.ac.uk/repository/bitstream/handle/10036/3689/daf_report_public.pdf?sequence=1

[40] Organization for Economic Co-operation and Development 2015. *Making Open Science A Reality*. https://www.innovationpolicyplatform.org/content/open-science

[41] Parsons, T. et al. 2013. *Research Data Management Survey*. University of Nottingham. http://admire.jiscinvolve.org/wp/files/2013/02/ADMIRe-Survey-Results-and-Analysis-2013.pdf

[42] Peng, G. et al. 2015. A Unified Framework for Measuring Stewardship Practices Applied to Digital Environmental Datasets. *Data Science Journal*. 13, 0 (Apr. 2015). http://datascience.codata.org/articles/abstract/10.2481/dsj.14-049/

[43] Pepe, A. et al. 2014. How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. *PLoS ONE*. 9, 8 (2014), e104798. http://dx.doi.org/10.1371/journal.pone.0104798

[44] Perry, C. 2008. Archiving of publicly funded research data: A survey of Canadian researchers. *Government Information Quarterly*. 25, 1 (2008), 133–148. http://www.sciencedirect.com/science/article/pii/S0740624X07000561

[45] Peters, C. and Dryden, A. 2011. Assessing the Academic Library's Role in Campus-Wide Research Data Management: A First Step at the University of Houston. *Science & Technology Libraries*. 30, 4 (2011), 387–403. http://dx.doi.org/10.1080/0194262X.2011.626340

[46] Pienta, A.M. et al. 2010. The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data. (Nov. 2010). http://deepblue.lib.umich.edu/handle/2027.42/78307

[47] Piwowar, H.A. and Chapman, W.W. 2008. Identifying Data Sharing in Biomedical Literature. *AMIA Annual Symposium Proceedings*. 2008, (2008), 596–600. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2655927/

[48] Podesta, J. et al. 2014. *Big Data: Seizing Opportunities, preserving values*. Executive Office of the President. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf

[49] Read, K.B. et al. 2015. Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study. *PLoS ONE*. 10, 7 (Jul. 2015), e0132735. http://dx.doi.org/10.1371%2Fjournal.pone.0132735

[50] Scaramozzino, J. et al. 2012. A Study of Faculty Data Curation Behaviors and Attitudes at a Teaching-Centered University. *College & Research Libraries*. 73, 4 (Jul. 2012), 349–365. http://crl.acrl.org/cgi/content/abstract/73/4/349

[51] Sunlight Foundation and Keserű, J. 2015. We're still looking for open data social impact stories! *Sunlight Foundation*. http://sunlightfoundation.com/blog/2015/02/25/were-still-looking-for-open-data-social-impact-stories/

[52] Sveinsdottir, T. et al. 2013. *Stakeholder values and relationships within open access and data dissemination and preservation ecosystems*. Policy RECommendations for Open access to research Data in Europe (RECODE). http://recodeproject.eu/wp-content/uploads/2013/10/RECODE_D1-Stakeholder-values-and-ecosystems_Sept2013.pdf

[53] Tenopir, C. et al. 2011. Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE*. 6, 6 (2011). http://dx.doi.org/10.1371/journal.pone.0021101

[54] Tenopir, C. et al. 2015. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLoS ONE*. 10, 8 (2015). http://dx.doi.org/10.1371/journal.pone.0134826

[55] The Royal Society 2012. *Science as an Open Enterprise*. http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE-Summary.pdf

[56] Thornhill, K. and Palmer, L. 2014. An Assessment of Doctoral Biomedical Student Research Data Management Needs. (2014). http://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1075&context=escience_symposium

[57] Turner, V. et al. 2014. *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. http://idcdocserv.com/1678

[58] Ubaldi, B. 2013. Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. *OECD Publishing*. No. 22 (May 2013). http://dx.doi.org/10.1787/5k46bj4f03s7-en

[59] UNC-CH 2012. *Research Data Stewardship at UNC: Recommendations for Scholarly Practice and Leadership*. University of North Carolina Chapel Hill. http://sils.unc.edu/sites/default/files/general/research/UNC_Research_Data_Stewardship_Report.pdf

[60] Vickery, G. 2011. *Review of Recent Studies on PSI Re-use and Related Market Developments*. http://www.dpconline.org/component/docman/doc_download/340-mind-the-gap-assessing-digital-preservation-needs-in-the-uk

[61] Waller, M. and Sharpe, R. 2006. *Mind the Gap: Assessing Digital Preservation Needs in the UK*. Digital Preservation Coalition. http://www.dpconline.org/component/docman/doc_download/340-mind-the-gap-assessing-digital-preservation-needs-in-the-uk

[62] Wallis, J.C. et al. 2013. If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLoS ONE*. 8, 7 (Jul. 2013), e67332. http://dx.doi.org/10.1371%2Fjournal.pone.0067332

[63] Wynholds, L. et al. 2011. When Use Cases Are Not Useful: Data Practices, Astronomy, and Digital Libraries. *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (New York, NY, USA, 2011), 383–386. http://escholarship.org/uc/item/4tk5d7hx

[64] York, J. et al. 2016. What Do We Know About The Stewardship Gap? https://deepblue.lib.umich.edu/handle/2027.42/122726

What Makes A Digital Steward:

A Competency Profile Based On The National Digital Stewardship Residencies

Karl-Rainer Blumenthal
Internet Archive
karlb@archive.org

Peggy Griesinger
George Mason University
mgriesin@gmu.edu

Julia Kim
Library of Congress
juliakim@loc.gov

Shira Peltzman
University of California, Los Angeles
speltzman@library.ucla.edu

Vicky Steeves
New York University
vicky.steeves@nyu.edu

ABSTRACT

Digital stewardship is the active and long-term management of digital objects towards their preservation for and unencumbered access by future generations. Although the field is rapidly maturing, it still lacks a comprehensive competency profile for practitioners. This is due in part to the relative youth of the field, and to the fact that being an effective steward of digital materials requires highly specialized training that is best acquired through hands-on work. Given the key role that competency profiles play in the design of curricula and job postings, the lack of one hinders the training and education of professionals for these positions. This paper provides a profile of the skills, responsibilities, and knowledge areas that define competency in digital stewardship, based on a close study of the projects undertaken in the National Digital Stewardship Residency program (NDSR). The authors use a triangulated research methodology in order to define the scope of the profile, qualitatively analyze the competencies articulated among NDSR project descriptions, and quantitatively evaluate those competencies' importance to professional success. The profile that results from this research has implications for current and future digital stewards: training designed with this profile as its basis will focus on the skills most needed to be an effective digital steward, and therefore can guide both graduate and professional development curricula alike.

Keywords

digital stewardship, National Digital Stewardship Residency, NDSR, education, training, digital preservation

1. INTRODUCTION

Although digital preservation is a young field, there are now more scholarship, tools, and resources that address the long-term stewardship[1] of digital material than ever before. In recent years there has been a notable expansion of educational and training resources in particular, including workshops, symposia, conferences, and professional development curricula. However, as the 2015 National Agenda for Digital Stewardship asserts, ”[g]enuine interest and motivation to learn about a subject cannot be taught in a workshop or training session; similarly, knowledge about standards and practices in an evolving field is best gained through direct, practical experience.” [1] In short, being an effective steward of digital material requires more extensive and specialized training than can be acquired through traditional means.

What, then, makes a digital steward? Despite the acknowledgment that stewards must possess a particular skillset, there has not yet been sufficient scholarship performed to identify a competency profile for digital stewards, as exists in other professional communities. A competency profile succinctly articulates the specific skills, responsibilities, and knowledge areas required to practice in one's profession, and is therefore instrumental to setting training and education goals. Perhaps it is due to the field's relative youth that so many analyses of it have focused principally on the surrounding literature–most commonly surveys of graduate school curricula or job advertisements–rather than on the backgrounds and training of practitioners themselves. But as the amount of digital material entering libraries, archives, and museums worldwide continues to grow, developing successful training goals for the next generation of stewards is an increasingly vital pursuit.

The lack of any cogent competency profile for this field is significant because competency profiles are used in the creation of job ads and curriculum development, which in turn affects how the field and its practitioners succeed in and improve their profession. In spite of this, the Agenda singles out the National Digital Stewardship Residency (NDSR hereafter) as an especially successful training model due to the fact that it allows recent graduates to gain practical, hands-on experience in the field managing digital stewardship projects. Although measuring the long-term impact of this program on the field at large would be premature[2], the project descriptions created by host institutions for both current and former residents yield valuable information. Both the wide variety of projects and activities covered as well as the fact that they explicitly outline goals and responsibilities for each individual resident and project makes them ideal for determining the skillset and expertise required to successfully perform the professional duties of a digital steward.

The authors developed a competency profile for digital stewards by using a three-pronged approach: 1) reviewing literature on the topics of emerging digital stewardship roles, responsibilities, expected practices, and training needs; 2) qualitatively analyzing current and completed NDSR project descriptions, which outline project tasks and deliverables; and 3) quantitatively analyzing the results from a survey conducted of former and current residents that identified the range and types of competencies required to successfully complete each project. The result is a profile of the skills, responsibilities, and knowledge areas that define competency in digital stewardship, which will create a clearer understanding of the on-the-job skills required of digital stewardship professionals in the hopes of informing future professional and curricula development in the field.

2. ABOUT NDSR

NDSR was created by the Library of Congress, in partnership with the Institute of Museum and Library Services (IMLS), with the mission to ”build a dedicated community of professionals who will advance our nation's capabilities in managing, preserving, and making accessible the digital record of human achievement.” [3] In its pilot year (2013-2014) NDSR matched ten recent graduates with mentors at ten cultural heritage institutions in order to develop, apply, and advance emerging digital stewardship practices and their own knowledge and skills in real-world settings. Since then, IMLS has granted funding to five additional NDSR programs among cultural heritage organizations throughout the country.

The program involves competitive selection processes for both host institutions and residents. Host institutions are selected on the basis of criteria such as their ability to provide higher-level support and mentorship to residents, as well as the significance of their proposed projects. These projects can be as broad in scope as institutional assessments and policy writing, or as narrow as documenting the particular application of software within a larger workflow. Applicants must be U.S. citizens or able to work in the U.S., as well as recent graduates of post-baccalaureate degrees. Although residents' salaries are paid through IMLS grant funds, they are regarded as regular employees by their host institutions and measures are taken to ensure that they are incorporated into the fabric of their institutions' workplaces. This is balanced by the fact that the residency is an apprenticeship program in which important criteria are learning outcomes and job placement within the field after its completion. Each NDSR program supplements on-site support with workshops and trainings designed to foster professional growth. Residents are also strongly encouraged to publicize their projects through presentations and conference participation.

3. LITERATURE REVIEW

Competency profiles are a common way for information management professions to express educational and/or professional benchmarks. These include foundational professional concepts, information resources, research standards, lifelong learning expectations, and management principles and ethics, among other things. The American Library Association's ”Core Competencies of Librarianship,” for instance, establishes a baseline for those things that every ”person graduating from an ALA-accredited master's program in library and information studies should know and, where appropriate, be able to employ.” [4] At least 16 affiliated or closely related professional organizations have adopted similar statements. [5]

Studies of training needs and efficacy [6–8] cite the lack of a commonly accepted profile for digital stewardship as confounding to efforts to design complementary curricula. Alternative approaches in the U.S., [9, 10], U.K. [11], and internationally [12, 13] survey professionals actively working in digital stewardship roles to identify their core competencies in order to broadly identify gaps and opportunities in the training and education of current and future professionals. Efforts continue to develop rigorous digital stewardship curricula among select ALA-accredited programs in library and information science. They range from exhaustively deductive matrices of technical proficiencies [14] to inductive and fieldwork-based practicum programs. [15]

Studies both external [16] and internal [17] to the Society of American Archivists (SAA) were instrumental to the creation of that organization's Digital Archives Specialist (DAS) Curriculum and its corresponding certification program, which at the time of writing provides the archival profession's most succinct, widely disseminated, and professionally supported profile of the ”core competencies” for digital archivists. These competencies are summarized in seven statements of ability, such as: ”#1. Understand the nature of records in electronic form, including the functions of various storage media, the nature of system dependence, and the effect on integrity of records over time.” [18] Digital stewards outside of the archives domain would benefit from similarly rigorous research and output.

The logic for identifying competency indicators differs across the above efforts, but the authors took especial interest in the methodology chosen for the *Information: Curate, Archive, Manage, Preserve* (iCAMP) curriculum development project, which reduces the language of data management job advertisements to summaries of the job titles, ex-

[1] For the purposes of this paper, ”digital stewardship” is defined as the active and long-term management of digital objects towards their robust preservation for and unencumbered access by future generations, inclusive of all subfields of labor and expertise previously defined among professional surveys and studies as digital curation, data curation, data management, digital archiving, digital preservation, and digitization. Digital stewards include data librarians, digital asset managers, digital archivists, and all manner of administrators who seek to align disparate digitization and digital preservation efforts.

[2] Although it is not a longitudinal analysis, the Council on Library and Information Resources (CLIR) is at the time of writing conducting a cross-cohort assessment of the entire NDSR program in order to evaluate the significance of the residency experience for the residents and their host institutions, and to identify common success factors across the various residencies. [2]

perience requirements, and knowledge and skill expectations that they contain [19]. The results are too specific to the data management domain and generalized in their language to answer this paper's questions regarding digital stewardship writ large. However, they provide a useful precedent for the application of qualitative data analysis tools to perform comparable document analysis on a corpus of residency project descriptions that the authors believe are both more broad in their professional scope and specific in their language.

Less rigorous, more impromptu investigations [20, 21] also mine the corpus of job advertisements for language articulating the specific competencies desired by information organizations hiring digital archivists. These inquiries provide useful insight into the emerging lexicon of digital archives, but leave open to question how many of these articulated competencies and skills are the core responsibilities for their hires, and towards which future professionals must train.

The literature review reveals an opportunity to provide digital stewards with an overarching competency profile and statement that span various specializations within the field, but which also articulate requirements concretely enough to guide graduate and professional education and training goals.

The authors used a triangulated approach to create a profile of digital stewardship competencies. The literature review provided an initial sample of commonly used summary terminology for skills, knowledge areas, and responsibilities typically applied in practice. This informed the authors' distillation of 35 NDSR project descriptions through document analysis[3], the results of which provided the authors the precise terminology with which to construct a survey instrument.

Project descriptions for both New York residency cohorts [23] and the second of the two cohorts in each Boston [24] and Washington, D.C. [25] were retrieved from each cohort's official website. Project descriptions for the initial Boston [26] and Washington, D.C. [27] residency cohorts were retrieved from the archived instances of those cohorts' official websites made available through the Internet Archive's Wayback Machine.

4. RESEARCH METHODS

The authors used a social science research methodology called grounded theory [28] to analyze the qualitative data (project descriptions). Research using grounded theory begins with a collection of qualitative data that the researcher then reviews and re-reviews. During this process, the researcher tags specific quotes, words, or phrases as evidence, and assigns them "codes" that represent larger ideas. [29] As data is iteratively reviewed, these codes can be grouped into concepts and ultimately categories, which become the basis for a new thesis or theory. This differs from traditional qualitative methodology because it creates its theoretical framework inductively, rather than relying upon an existing one. [30]

The authors used this method to code for attributes expected of each resident. In order to do this, the authors used NVivo[4], a proprietary qualitative data analysis software designed for researchers working with data that requires deep levels of analysis. NVivo was chosen because of its real-time version control, which was useful because the research team was geographically distributed. Two of the authors performed an initial blind review of the materials, using a predetermined codebook[5] based on an initial sampling of the dataset and the literature review.

Although the document analysis could provide the authors with a baseline understanding of the attributes that the residents were intended to develop, the authors also sought to examine how the projects had been borne out in practice. To accomplish this, the authors designed and implemented an online survey of current and past residents. By comparing the findings of the document analysis and the survey, the authors could assign quantitative weight to any similarities, differences, or unanticipated but necessary competencies.

The overarching code categories became the question blocks and the sub-codes became the corresponding rating matrix of individual questions within the survey instrument (see Supplementary Materials). The authors chose to exclude *Personality requirements* (see Table 1) from the survey because these are general traits common to job advertisements across professions, rather than specific to digital stewardship.

The authors used Qualtrics[6], a proprietary research software used to enable online data collection through building survey instruments, because it was readily available via an institutional license, randomized question order, and anonymized participants down to the IP address.

Initially, four survey invitations were sent to the list of participants using the Qualtrics email function, or "mailer." The mailer allows for complete anonymity in the data collection: the authors could not see who had completed or not completed the survey. This also allowed the authors to send out individualized, anonymous links, to separate respondents in bulk. Nine current or former residents did not participate by the date on which the survey was originally scheduled to end. To get as close to a full dataset as possible, each author sent a follow-up email to four-to-seven participants to remind them of the deadline. The link to the survey included in these emails was still anonymous and did not record IP addresses, but was no longer unique to each recipient.

Table 1: **Code categories, their frequencies and sub-codes from the document analysis.**

Code category	Frequency	Sub-Codes
Technical skills	397	Format migration/transcoding Metadata Workflow enhancement/development Audio/video Digital asset management Digitization Coding/scripting Implementation of hardware/software Web archiving Qualitative and data analysis skills
Professional output responsibilities	275	Metadata crosswalk/guidelines Report/recommendations Survey/inventory Teaching materials/toolkits Scholarly output
Communication skills	148	Presentation Written output Workshop/training Interact/liaise with internal staff/stakeholders Interact/liaise with external stakeholders Public outreach
Research responsibilities	118	Literature review Survey of standards/best practices Environmental scan
Project management abilities	92	Managing resources Managing people
Knowledge of standards and best practices	62	Metadata Data management Repository management
Personality requirements	30	Attention to detail Flexible Enthusiastic

The authors acknowledge several methodological issues with the data collection for this study. The first is that the authors are included in the dataset as participants. The most significant issue is that the authors effectively studied themselves; they designed, tested, and discussed the survey before deployment. As a result, they did not take the survey blind. Not only did this differentiate them from the rest of the participants, which could potentially skew the data, but it also introduced the potential for nonresponse bias [31]. However, the authors randomized the questions to mitigate the latter issue. Although the authors recognize that participating in their own research is unorthodox, they felt that it was essential to equally represent all of the different NDSR projects, locations, and cohorts in the survey results were they to recuse themselves. Moreover, because the authors all belonged to the same 2014-15 NDSR in New York cohort, those projects would not have been represented in the survey results. The authors felt that the benefits of including their responses outweighed the potential costs of excluding their responses from the dataset.

Another potential problem was the fact that fifteen of the participants took the survey before they completed their residencies. This introduced a possibility for survey bias [32]. They might not have been able to answer the optional questions regarding 1) post-NDSR job functions, and 2) additional skills necessary to complete their residencies. However, since the current residents could answer all the required questions (they were more than halfway through their residencies during data collection), they were still included in the participant population.

The authors' final concern was with sending individual emails to participants. This demystified some of the initial anonymity afforded by using the Qualtrics mailer. Some participants replied to these individualized emails, indicating they had already taken the survey (some even providing the date), or that they had not taken part but would do so shortly. The authors promptly deleted these emails permanently, and no records of them remain. Given the already small sample size, the authors felt that having as close to a complete dataset as possible was so impactful to the results that the follow-ups were necessary.

5. RESULTS

This study had two main outputs: the results of the document analysis (qualitative), and the results of the survey (quantitative). Through examining both, the authors could create a matrix of the competency areas vital to the National Digital Stewardship Residencies.

5.1 Document Analysis

Two of the authors coded the project descriptions. In order to compare their interpretations of the data, the authors used the NVivo "coding comparison" feature to determine that they had a 90% agreement rate on the codes, and then met to reconcile the 10% of cases in which their coding differed. The seven resulting high-level code categories

[3] Document analysis is a systematic procedure for analyzing and interpreting data generated from documents; in qualitative research, document analysis is often used to corroborate findings from other data sources such as surveys, interviews, etc. [22]

[4] Produced by QSR International: http://www.qsrinternational.com/product

[5] A codebook describes and defines the codes for which the authors searched.

[6] Produced by Qualtrics: https://www.qualtrics.com/

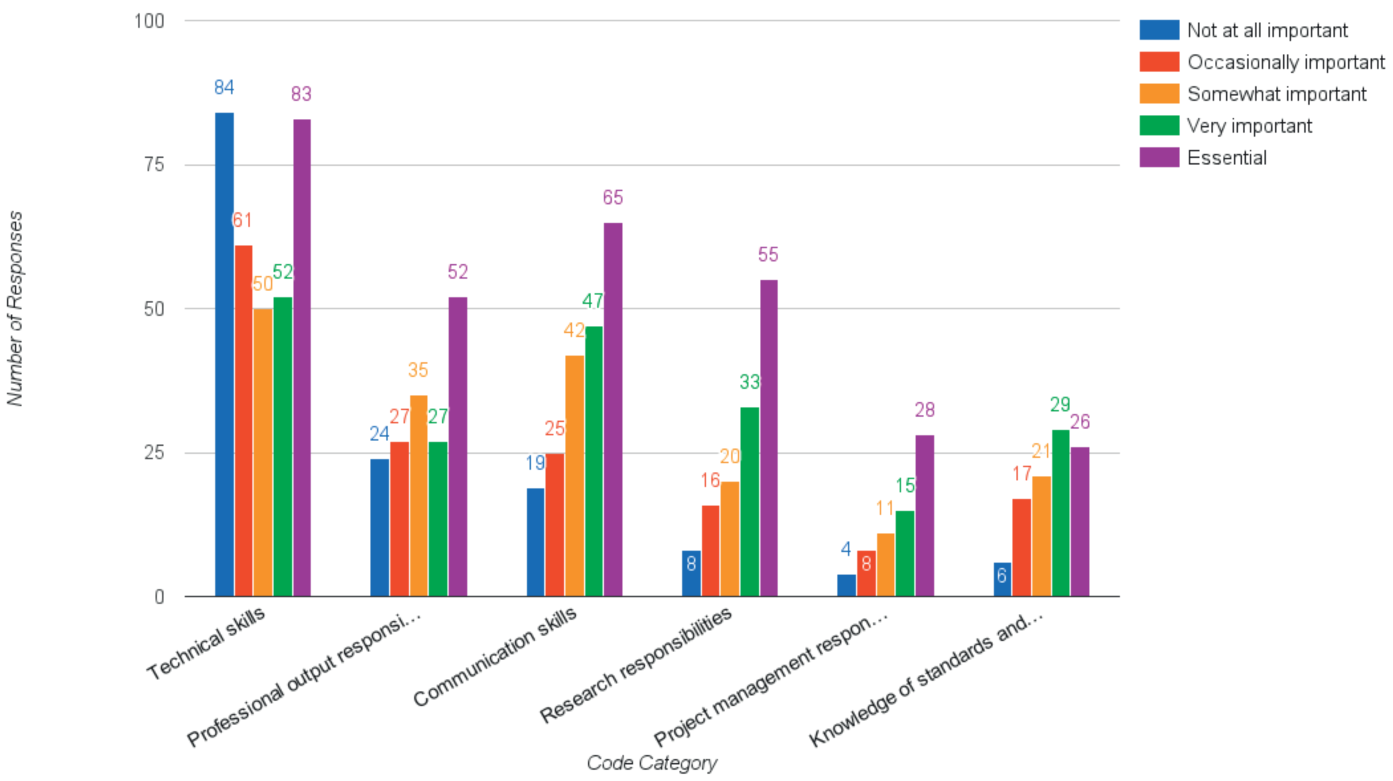


Figure 1: **Total distribution of frequency of responses over code categories.**

represent the overall categories of competencies required to perform as a digital steward. These were informed by terminology from the literature review and the initial sampling of the qualitative dataset.

Seven coded categories of competence in residency-related functions emerged from the analysis: *Technical skills*; *Knowledge of standards and best practices*; *Research responsibilities*; *Communication skills*; *Project management abilities*; *Professional output responsibilities*; and *Personality requirements*. The authors iteratively reviewed the qualitative data in order to identify sub-codes that more specifically represent the competency areas applied in the performance of the residencies. The minimum number of sub-codes per category was two, within *Project management abilities*, and the maximum was ten, within *Technical skills* (see Table 1).

Due in part to their extensive range of skills, *Technical skills* has the highest frequency of appearances in the data (397). The second-highest is *Professional output responsibilities* (275). *Personality requirements* appear the least, at 30 in total.

5.2 Survey Responses

The survey was open from March 14 to April 1, 2016. After excluding *Personality requirements*, each of the six code categories had one required question, which took the form of a rating matrix (see Supplementary Materials). Each sub-code (see Table 1) represented a row of the matrix, and participants were asked to rank competencies on a five-point Likert scale from "Not at all important" (1) to "Essential" (5) (matrix columns). The survey had a 94% response rate, having received 33 participant responses out of a total group of 35. The authors analyzed the frequency of responses in each code block in order to determine what the most impactful categories and individual competencies were to achieving residency goals.

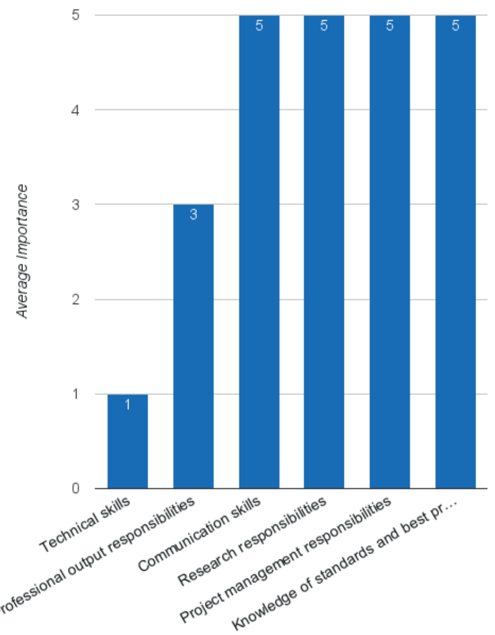


Figure 2: **Modal average of code categories.**

Table 2: **Responses per sub-code with descriptive statistics.**

		Response Counts					
Code Category	**Sub-Code**	**1**	**2**	**3**	**4**	**5**	**Mode**
Technical skills	Format migration/transcoding	10	6	10	4	3	1,3
	Metadata creation and manipulation	5	5	5	7	11	5
	Workflow enhancement/development	1	1	3	7	21	5
	A/V preservation	14	6	3	2	8	1
	Digital asset management	2	3	5	8	15	5
	Digitization	11	8	5	5	4	1
	Coding/scripting	10	10	6	4	3	1,2
	Hardware/software implementation	6	7	7	7	6	3
	Web archiving	18	7	3	1	4	1
	Qualitative data analysis	7	8	3	7	8	2,5
Professional output responsibilities	Metadata documentation	6	4	7	9	7	3,5
	Reports/recommendations	0	3	0	2	28	5
	Surveys and/or inventories	1	9	8	7	8	2
	Teaching materials/toolkits	7	7	10	3	6	3
	Scholarly output (ie. annotated bibliographies, white papers, etc.)	10	4	10	6	3	1,3
Communication skills	Presentations (webinars, conferences, in-person stakeholder meetings, etc.)	1	2	6	9	15	5
	Written output (blog posts, journal articles, etc.)	2	1	14	9	7	3
	Workshops and trainings	3	8	9	7	6	3
	Internal Interactions	0	0	1	8	24	5
	External Interactions	4	5	6	8	10	5
	Public Outreach (social media, public events, etc.)	9	9	6	6	3	1,2
Research responsibilities	Literature reviews	6	7	9	5	6	3
	Surveys of best practices and standards	0	5	4	7	17	5
	Environmental scans (e.g. reviewing practices at peer institutions)	1	3	2	11	16	5
	Needs assessment/gap analysis	1	1	5	10	16	5
Project management abilities	Managing project resources (ie. workflows, tools, documentation, etc.)	2	2	2	9	18	5
	Managing people (ie. vendor relations, intern/staff supervision, etc.)	2	6	9	6	10	5
Knowledge of standards and best practices	Metadata	1	5	5	14	8	4
	Data management	4	5	7	8	9	5
	Repository management (TRAC, TRD, OAIS, etc.)	1	7	9	7	9	3,5

Respondents identified "Essential" competencies frequently throughout the code categories. In four of the six categories, more sub-codes were deemed "Essential" than they were deemed any other level of importance, and in each case by a margin of at least seven responses. *Technical skills* had a higher combined frequency of responses for "Not at all important" and "Occasionally important" than it did for "Very important" and "Essential," which drove down its overall importance rating. Only three more respondents deemed *Knowledge of standards and best practices* "Very important" than those who deemed it "Essential."

The category of *Technical skills* had the lowest average importance of the six codes, and *Professional output responsibilities* had the second-lowest. The other four codes were all deemed "Essential" on average by the participants (see Figure 3).

5.2.1 *Technical Skills*

Perhaps the most striking aspect of the data was the *Technical skills* category. *Technical skills* had the most mixed results of any category in the survey, which could be due in part to the fact that it had the highest number of granular competency areas (sub-codes). The result was a clear disparity in the distribution of responses per importance level. The outlier in *Technical skills* with the lowest importance rating was *Web archiving*, which lowered the overall importance found in Figure 3. *Workflow enhancement* was also an outlier; it was rated as the most essential technical skill by a margin of seven responses.

5.3 Optional Questions

After answering the required questions above, survey respondents were invited to answer three optional questions.

5.3.1 *Quantitative*

An optional question in the survey asked the participants whether or not their experience in NDSR was relevant to their current employment. Every participant answered this question, with 90% (30 participants) say yes, while 10% (3 participants) answering no.

5.3.2 *Qualitative*

The last two questions in the survey were open-ended questions that asked participants for feedback in longer-form writing. The first question asked participants to identify any competencies not addressed in the survey. 33% (11 of 33) of respondents answered this question. The authors could not ascribe any particular pattern to these responses, however several of them further described a competency or competencies from the survey as applied to their specific project. The second question asked for any additional feedback or comments. 18% (6 of 33) answered the second question. These answers were not analyzed using the qualitative methods above due to the low frequency and disparate topics covered, some of which again answered the previous optional question.

6. CONCLUSIONS

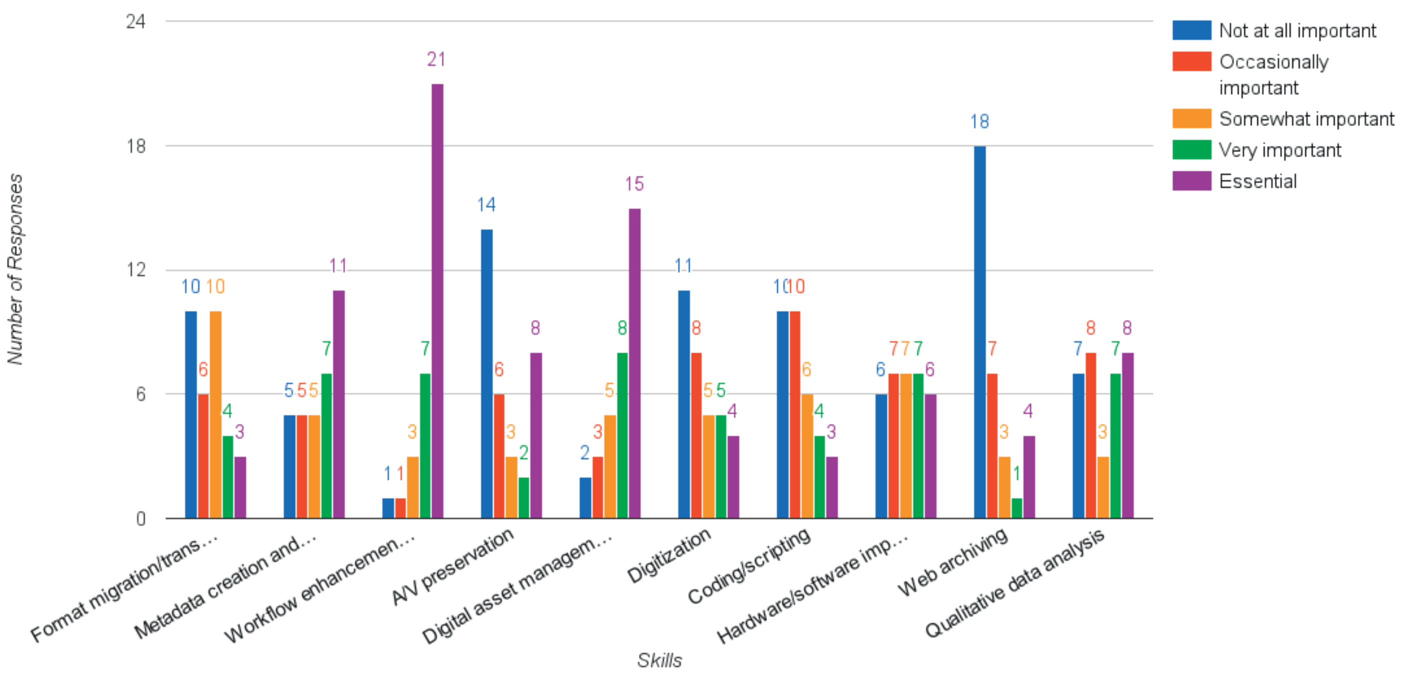


Figure 3: **Breakdown of technical skills code category.**

Learning from the competency areas that were described in the NDSR projects and identified by residents as being especially important (i.e. achieving a surveyed modal average of 4 [Very important] or 5 [Essential]), a competency statement representing this profile could read as follows:

> Effective digital stewards leverage their technical skills, knowledge of standards and best practices, research opportunities, communication skills, and project management abilities to ensure the long-term viability of the digital record.
>
> In order to accomplish this, they cultivate their skill developing and enhancing new and existing digital media workflows, managing digital assets, and creating and manipulating these assets' metadata. They commit to the successful implementation of these new workflows by reliably managing both project resources and people.
>
> They maximize the impact of their professional practice by soliciting regular input from stakeholders both internal and external to their institutional setting. They articulate and document the standards and practices that address these needs by creating policies, professional recommendations, and reports, which requires that they maintain current and and expert knowledge of standards and best practices for metadata and data management in their respective sectors.
>
> They articulate and document the practices that address these needs by creating policies, professional recommendations, and reports, which requires that they maintain current and expert knowledge of metadata and data management standards in their respective sectors.
>
> Digital stewards are qualified to manage, preserve, and provide access to various new and/or challenging forms of media. They may also engage in, among other things: coding and scripting; digitization; hardware and software implementation; public outreach; and special media format management and migration.

The authors conclude that while there are some fundamental competencies required of digital stewards, digital stewardship also encompasses niche skills that are role-specific. Several *Technical skills* were far more important to some projects than to others, and therefore could be considered specialized, rather than fundamental skills. There was a clear bimodal distribution for *Technical skills* (sub-codes in this category were deemed "Not at all important" 84 times and "Essential" 85 times). The authors posit that while job postings often list *Technical skills* as being essential, this study indicates that they are not always essential to all jobs in practice.

These split distributions apply to *Technical skills* sub-codes as well. For example, respondents were evenly split when gauging the importance of both *Hardware/software implementation* and *Qualitative data analysis*. These skills were unambiguously important to half of the respondents, but unambiguously unimportant to the other half. *Web archiving* distinguishes itself in this regard as a particularly niche skill–"Essential" to four respondents, but "Not important at all" to eighteen. By contrast, *Workflow enhancement* is a universally important skill, having been deemed "Essential" twenty-one times and "Not important at all" only once.

By analyzing the project descriptions of the National Digital Stewardship Residencies, the authors enumerated the competency areas that define digital stewardship across a broad swath of applications. By surveying the residents responsible for successfully completing these residencies, they were also able to highlight fundamental competency areas that therefore belong in any profile of an effective digital steward.

7. IMPLICATIONS AND FUTURE WORK

While the majority of competencies (sub-codes) surveyed for this study were definitively fundamental (had a mode ≥ 4) or specialized (had a mode of ≤ 2), there were thirteen that could not be as conclusively categorized. Of these, there were five that had a mode of 3, meaning the majority of the participants labeled these as "Somewhat important." These are: *Hardware/software implementation*, *Written output*, *Workshops and trainings*, *Teaching materials/toolkits*, and *Literature review*. Seven sub-codes had multiple modes, showing disagreement among the participants as to the relevance of the skill for successfully completing their digital stewardship work. These are: *Format migration/transcoding*, *Coding/scripting*, *Qualitative data analysis*, *Public outreach*, *Repository management*, *Metadata documentation*, and *Scholarly output*. The authors refrained from assigning these sub-codes into either the "Fundamental" or the "Specialized" tiers. The authors included them in this study's resulting competency statement as examples of further and increasingly specialized areas of work for which digital stewards are qualified, however, determining the place that these specific thirteen sub-codes hold in the overall profile of competencies for digital stewards presents an opportunity for future research.

It is important to note that this study's qualitative analysis was based on descriptions of projects, all of which were inherently time-limited and some of which were deliberately narrow in focus. While it was beyond the scope of this study, the diversity of project types among NDSR cohorts may also have affected the results. The specificity of certain projects, coupled with the fact that they were all designed to be accomplished in a relatively short time-frame, may have impacted our results to some degree–perhaps enough so to merit a new study that is based on a different set of data. However, the 90% affirmation among this study's survey respondents implies that these competencies extend to digital stewardship positions beyond NDSR. The authors encourage using a similarly triangulated methodology to analyze competency areas found among permanent position descriptions and their incumbents. In particular, a follow-up study of those who have completed National Digital Stewardship Residencies and are now in permanent digital stewardship positions could do so while counterbalancing any possible bias of this study towards competencies that apply disproportionately to short-term appointments.

Finally, it is worth noting the fact that all residencies took place in the U.S.A., and consequently that this research is not international in scope. This presents an important area for future research, which might involve conducting a comparable study built on job descriptions culled from a variety of national contexts. Contrasting the results of such a study with the competency profile presented here would perhaps enable the construction of a stronger and more well-rounded profile overall.

This research has implications for current and future digital stewards alike: The resulting profile can be used to guide graduate and professional development curricula, and training designed with this profile as its basis will focus on the skills most needed to be an effective digital steward. For instance, this study suggests that although specific technical skills are viewed as highly important in different settings, a much larger majority of projects required skills less bound to a particular technology or media, like documentation creation and workflow analysis. The high level of agreement regarding the importance of writing reports and communicating internally also bolsters a need for digital stewards to not only possess a deep understanding of their field, but to effectively disseminate their work to others. This new profile illustrates the fundamental competencies that must be cultivated by digital stewards in order to succeed in the profession.

8. SUPPLEMENTARY MATERIALS

The authors welcome and encourage others to extend and reproduce their study, and have made all research materials, including the survey instrument and data, freely available at the following URL: `https://osf.io/xfc26`

9. REFERENCES

[1] National Digital Stewardship Alliance. National Agenda for Digital Stewardship. September 2015.

[2] Council on Library and Information Resources. CLIR Receives IMLS Grant to Assess National Digital Stewardship Residency Programs. http://web.archive.org/web/20160313223145/http://www.digitalpreservation.gov/documents/2015NationalAgenda.pdf, September 2015.

[3] Library of Congress. National Digital Stewardship Residency. http://web.archive.org/web/20160404223713/http://www.digitalpreservation.gov/ndsr/, 2013.

[4] American Library Association. ALA's Core Competences of Librarianship. http://web.archive.org/web/20160304200942/http://www.ala.org/educationcareers/sites/ala.org.educationcareers/files/content/careers/corecomp/corecompetences/finalcorecompstat09.pdf, 2009.

[5] American Library Association. Knowledge and competencies statements developed by relevant professional organizations. http://web.archive.org/web/20150830202610/http://www.ala.org/educationcareers/careers/corecomp/corecompspecial/knowledgecompetencies, 2015.

[6] W.M. Duff, A. Marshall, C. Limkilde, and M. van Ballegooie. Digital Preservation Education: Educating or Networking? *The American Archivist*, 69(1):188–212, 2006.

[7] H.R. Tibbo. So much to learn, so little time to learn it: North American archival education programs in the information age and the role for certificate programs. *Archival Science*, 6:231–245, 2006.

[8] C. Thomas and S.I. Patel. Competency-based training: a viable strategy for an evolving workforce? *Journal of Education for Library and Information Science*, 49(4):298–309, 2008.

[9] Y. Choi and E. Rasmussen. What is needed to educate future digital librarians: a study of current practice and staffing patterns in academic and research libraries. *D-Lib Magazine*, 12(9), 2006.

[10] Cal Lee and Helen Tibbo. Closing the Digital Curation Gap Focus Groups Report, June 2011.

[11] G. Pryor and M. Donnelly. Skilling up to data: Whose role, whose responsibility, whose career. *International Journal of Digital Curation*, 2(4):158–170, 2009.

[12] Ragnar Andreas Audunson and Nafiz Zaman Shuva. Digital Library Education in Europe. *SAGE Open*, 6(1):2158244015622538, January 2016.

[13] M. Madrid. A study of digital curator competences: a survey of experts. *International Information & Library Review*, 45(3-4):149–156, 2013.

[14] C. Lee. Matrix of Digital Curation Knowledge and Competencies. `http://web.archive.org/web/20150929215051/http://www.ils.unc.edu/digccurr/digccurr-matrix.html`, 2009.

[15] B. Fulton, P. Botticelli, and J. Bradley. DigIn: a hands-on approach to a digital curation curriculum for professional development. *Journal of Education for Library and Information Science*, 52(2):95–109, 2011.

[16] Digital Preservation Outreach and Education (DPOE). DPOE training needs assessment survey, 2010. http://web.archive.org/web/20160323170209.

[17] H.R. Tibbo. Presidential Address: On the occasion of SAA's diamond jubilee: A professional coming of age in the digital era. *American Archivist*, 75:17–34, 2012.

[18] Society of American Archivists. DAS Curriculum Structure. `http://web.archive.org/web/20150828210835/http://www2.archivists.org/prof-education/das/curriculum-structure`, August 2015.

[19] J. Kim, E. Warga, and W. Moen. Digital curation in the academic library job market. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–4, 2012.

[20] P. Chan. What Does it Take to Be a Well-rounded Digital Archivist? | The Signal: Digital Preservation. `http://web.archive.org/web/20160301155958/http://blogs.loc.gov/digitalpreservation/2014/10/what-does-it-take-to-be-a-well-rounded-digital-archivist/`, October 2014.

[21] J. Dooley. What's in a digital archivist's skill set? `http://web.archive.org/web/20150925105555/http://hangingtogether.org/?p=3912`, June 2014.

[22] G.A. Bowen. Document analysis as a qualitative research method. *Qualitative Research Journal*, 9(2):27–40, 2009.

[23] NDSR-NY. The National Digital Stewardship Residency in New York. `http://ndsr.nycdigital.org/`, 2015.

[24] NDSR-Boston. The National Digital Stewardship Residency Boston, Massachusetts. `http://projects.iq.harvard.edu/ndsr_boston/home`, 2016.

[25] NDSR. The National Digital Stewardship Residency. `http://www.digitalpreservation.gov/ndsr/index.html`, 2015.

[26] NDSR. The National Digital Stewardship Residency. `http://web.archive.org/web/20140801145633/http://www.digitalpreservation.gov/ndsr/index.html`, 2014.

[27] NDSR-Boston. The National Digital Stewardship Residency Boston, Massachusetts. `http://web.archive.org/web/20150320184022/http://projects.iq.harvard.edu/ndsr_boston/home`, 2015.

[28] K. Charmaz. *Constructing Grounded Theory*. SAGE Publications Inc, London, 2014.

[29] S. Sarker, F. Lau, and S. Sahay. Using an adapted grounded theory approach for inductive theory building about virtual team development. *Database for Advances in Information Systems*, 32(1):38–56, 2001.

[30] J.D. Orton. From inductive to iterative grounded theory: Zipping the gap between process theory and process data. *Scandinavian Journal of Management*, 13(4):413–438, 1997.

[31] M. Denscombe. *The good research guide: for small-scale social research projects*. Open University Press, Maidenhead, 5. ed edition, 2014.

[32] F. J. Fowler. *Survey research methods*. Applied social research methods series. SAGE, Los Angeles, fifth edition edition, 2014.

Identifying Barriers To File Rendering In Bit-level Preservation Repositories: A Preliminary Approach

Kyle R. Rimkus
University Library
University of Illinois at Urbana-Champaign
rimkus [at] illinois [dot] edu

Scott D. Witmer
School of Information Sciences
University of Illinois at Urbana-Champaign
sdwitme2 [at] illinois [dot] edu

ABSTRACT

This paper seeks to advance digital preservation theory and practice by presenting an evidence-based model for identifying barriers to digital content rendering within a bit-level preservation repository. It details the results of an experiment at the University of Illinois at Urbana-Champaign library, where the authors procured a random sample of files from their institution's digital preservation repository and tested their ability to open said files using software specified in local policies. This sampling regime furnished a preliminary portrait of local file rendering challenges, and thus preservation risk, grounded not in nominal preferences for one format's characteristics over another, but in empirical evidence of what types of files present genuine barriers to staff and patron access. This research produced meaningful diagnostic data to inform file format policymaking for the repository.

Keywords

digital preservation; file format policy; random sampling

1. INTRODUCTION

File formats are important to digital preservation—but are they understood? Repository managers often require or recommend specific formats over others, believing that favored file varieties will give their digital content a better chance at long-term viability than the riskier alternatives. This practice comes with acknowledged limitations. As DeVorsey and McKinney explain, "…files contain multifarious properties. These are based on the world of possibilities that the format standard describes, but can also include non-standard properties. The range of possibilities and relationships between them is such that it is quite meaningless to purely measure a file's adherence to a format standard" [4]. In other words, one ought to take endorsements of file formats in name only with a grain of salt, in lieu of better methods for representing the technical conditions necessary for the accurate rendering of digital content. This problem is explored in the Literature Review below, and is at the heart of the experiment presented in this paper.

2. LITERATURE REVIEW

As a young field, digital preservation is short on empirical evidence of file format risk, and most literature on the subject has been speculative in nature. In their 1996 report *Preserving Digital Information*, Waters and Garret suggested that repository managers faced with curating massive collections might adopt the practice of normalizing sets of heterogeneous file types to a smaller number of trusted formats [17]. Subsequently, repository managers and digital preservation researchers sought consensus on this approach, striving in particular to learn what qualities distinguish a trustworthy file format from an untrustworthy one.

Numerous studies, e.g., work conducted at the National Library of the Netherlands [12], Stanford University [1], and the Online Computer Library Center [15], strove to identify risk factors inherent to file formats. These research efforts, while complemented by the dissemination of public file format recommendations by institutional repository managers [11], have not however led to consensus on what qualities make a file format unassailably good. For example, many practitioners favor open over proprietary file formats because the way they encode content is transparent and publicly documented. On the other hand, the broad adoption of a proprietary file format by an active user community tends to ensure ongoing software support, and therefore long-term accessibility, for the format in question. Thus, it isn't always clear whether a particular external factor will without doubt positively or negatively affect a file format's long-term viability.

Becker et al point out that the "passive preservation" of bit-streams, even in so-called trusted file formats, is most effective when complemented by permanent access to legacy software environments [2]. This point of view has been elaborated by David Rosenthal, who challenges the utility of file format risk assessment, emphasizing that genuinely endangered formats are often so obscure or proprietary that no known rendering software exists for them in contemporary operating systems. In such cases, Rosenthal advocates for bit-level preservation of endangered files along with their fully emulated rendering environments [13].

Recent research has encouraged a situational approach to managing file format risk in repositories. In her 2014 paper "Occam's Razor and File Format Endangerment Factors," Heather Ryan denigrates the term file format *obsolescence* in favor of *endangerment* "to describe the possibility that information stored in a particular file format will not be interpretable or renderable using standard methods within a certain timeframe" [14]. This line of thinking is shared by a British Library study of that same year which posits that academic fretting over whether file format obsolescence exists or not is irrelevant in practice: "Working on the assumption that data in the vast majority of file formats will be readable with some degree of effort does not take into account two crucial issues. Firstly, what is the degree of effort to enable rendering, and what does it mean for an organization...?" [8]. Or, as DeVorsey and McKinney point out, risk assessment policies tend to stress the evaluation of potential external threats to digital files rather than the properties of the formats themselves: "At risk is not an inherent state of files and formats, it is an institution's view of its content determined by the policies, guidelines, and drivers it has at any one point in time" [4].

In a 2013 publication, an author of the present study found that the digital preservation file format policies of Association of Research Library member institutions were "very much rooted in relatively small-scale data management practices—stewarding files through digitization workflows, for example, or curating a university's research publications," but that, "As libraries and archives begin to set their sights on collections of heterogeneous files such as born-digital electronic records and research data, this is expected to spur on further evolution not only in the file formats that appear in digital preservation policies, but in the way file format policies are articulated and implemented" [11].

There is however a dearth of studies investigating the capacity of organizations to identify and assess file format risk as it exists within their repositories. Holden conducted a 2012 sampling and analysis of files on archived web pages conducted at France's Institut national de l'audiovisuel [5]. Similarly, Cochran published a report on file rendering challenges faced by the National Library of New Zealand [3]. In a similar vein, and influenced by concepts of organizational file format endangerment elaborated above, this paper seeks an evidence-based approach to assessing challenges to file rendering in bit-level preservation repositories.

3. BACKGROUND

In 2012, the University of Illinois at Urbana-Champaign (hereafter Illinois) Library established the Medusa digital preservation repository[1] for the long-term retention and accessibility of its digital collections. These consist primarily of digitized and "born digital" books, manuscripts, photographs, audiovisual materials, scholarly publications, and research data from the library's special collections, general collections, and institutional repositories. All master files created by the library's digitization units, for example, are by default deposited into Medusa.

Developed and managed locally by the Illinois library's repository group[2], Medusa features a web-accessible management interface, which provides collection managers with tools for initiating preservation actions. It provides forms for editing collection-level descriptive, administrative, and rights metadata; allows for the download of files or batches of files; tracks preservation events, file provenance, and file statistics; and provides on-demand verification of file fixity (md5 checksum values) and the extraction of technical metadata using the File Information Tool Set[3] (FITS) for files or groups of files. The library manages Medusa file storage in partnership with the National Center for Supercomputing Applications, also located on the Illinois campus. Medusa's storage infrastructure consists of two copies of every file replicated daily across two distinct campus nodes, both on spinning disk, and a third copy of every file backed up and stored out of state on magnetic tape.

As of March 23, 2016, the Medusa repository houses 8,209,807 files requiring just over 60 terabytes of storage space (180 if one takes into account all three copies). These files are predominately in image formats, but also feature a significant number of text, audio, and video files, also in a variety of formats.

The variegated nature of digital content housed in Medusa stems from the many departmental libraries, special collections units, scholarly communication initiatives, and grant-funded digitization projects the repository serves. Its collections derive however from five key areas of focus. The first three of these, which began in earnest in 2007, are: 1) the largescale digitization of books, newspapers, and documents, both in-house and in partnership with external vendors; 2) the digitization of special collections manuscript content conducted on-site or with vendors; and 3) the deposit of scholarly publications and other materials related to teaching and learning into the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS)[4] institutional repository. The other two areas of focus, which began gathering momentum in 2012, are: 4) the acquisition of born digital electronic records in the University Archives, and 5) the digitization of audio and moving image content from the special collections undertaken on site or by vendors (see Table 1).

Table 1. Approximate distribution of content source in Medusa repository by size

Source	Size (TB)
Digitized books, newspapers, documents	39
Digitized manuscripts, photographs, maps	10
Digitized audio and video	8
Born digital electronic records	2
Institutional repository (self-deposit)	1
TOTAL	**60**

Medusa does not at present enforce file format validation or normalization on ingest. While Medusa managers acknowledge these as best practices, they have sought, in their initial phase of provisioning a preservation repository, to focus on collection-level control of their holdings, stable storage, and bit-level services such as fixity monitoring and file format identification. Prior to the existence of the Medusa digital preservation service, collection curators at Illinois had stored archival master files on a variety of storage media, many of them precarious. These included optical disks, portable hard drives, and file servers without consistent backup. Having taken custody of more than 8,000,000 files in Medusa's first four years of existence, its managers are now interested in answering the following question: What are the most prevalent barriers to file access for curators and patrons who try to open files in Medusa's collections?

4. METHODOLOGY

4.1 Medusa Feature Development

According to specifications provided by the authors, developer Howard Ding introduced three new features in the Medusa web application to enable data collection and analysis:

1. Testing Profiles
2. Random Sampler
3. File Tester

4.1.1 Testing Profiles

The authors created a Testing Profile[5] to specify rendering conditions for each file format tested. Every Testing Profile listed a particular set of known extensions and MIME type values for a given file format. In addition, it specified the software, software version, operating system environment, and operating system version the authors would use for testing.

In identifying operating system and software values, the authors gave preference to tools deployed on site for library staff and users. Illinois Library Information Technology presently supports the Windows operating system for the majority of its employees, and web logs show that most library patrons also use Windows to access library resources. During the testing period, the operating system version of choice—for library staff and many patrons, and thus for this experiment—was Windows 7. The research goal being to assess file format challenges within the local access environment, this ensured results of practical relevance to collection curators and the communities they serve.

[1] https://medusa.library.illinois.edu/

[2] Source code for the Medusa collection registry application and its integrated microservices is available on Github (https://github.com/medusa-project).

[3] http://projects.iq.harvard.edu/fits/home

[4] https://www.ideals.illinois.edu/

[5] Go to https://medusa.library.illinois.edu/file_format_profiles for a full list of current profiles.

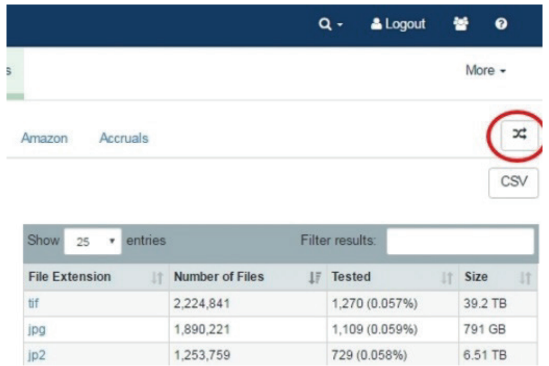
As an example, the profile for the format "TIFF" reads:

TESTING PROFILE: TIFF
Software: Adobe Photoshop
Software Version: CC2015
OS Environment: Windows
OS Version: 7
MIME types: image/tiff
File Extensions: tif, tiff

The authors emphasize that their approach to defining "file formats" in relation to these Testing Profiles constitutes a shorthand, and that the format standards under analysis can frequently take many forms. However, the use of such shorthand was deemed suitable to the purpose of this study.

4.1.2 Random Sampler

The Random Sampler provided the authors, at the click of a button, a file selected randomly from the repository for testing.



File Extension	Number of Files	Tested	Size
tif	2,224,841	1,270 (0.057%)	39.2 TB
jpg	1,890,221	1,109 (0.059%)	791 GB
jp2	1,253,759	729 (0.058%)	6.51 TB

Figure 1. Medusa dashboard file statistics view (Random Sampler button circled in red)

4.1.3 File Tester

The File Tester provides an interface for logging the success or failure of attempts to open files according to Testing Profiles. Specifically, it logs the operator, the date of the test, the Testing Profile in use, whether the test passed or failed, notes pertinent to the examination, and, in the case of failure, the reason why.

4.2 Testing Steps

The authors followed the steps below to gather data for this study:

1. Navigate to Medusa "dashboard" and press Random Sampler button (Figure 1)
2. Run technical metadata extraction tool File Information Tool Set (FITS)[6] on randomly selected file
3. Download and open file according to its corresponding Testing Profile
4. Fill out Analysis form with results of test (Pass/Fail, with reason for failure logged)

The authors assigned the status "Pass" to files that opened in the software program specified by their format profile without apparent rendering problems. If problems were apparent, they assigned the status "Fail," and appended a reason for the failure to the test record.

A sample test result reads:

FILE TEST: 00000004.jp2
UUID: 714621f0-5cb8-0132-3334-0050569601ca-f
Tester Email: email@illinois.edu
Date: 2015-12-08
Testing Profile: JPEG2000
Status: Fail
Notes: Renders in Kakadu, but not in Photoshop.
Test Failure Reasons: Software's file format module cannot parse the file

4.2.1 Constraints on Pass/Fail Criteria

Given the "multifarious" properties of computer files, a binary pass/fail distinction when evaluating files is no simple proposition. For this reason, the authors placed constraints on evaluations for several types of files:

- Files that clearly required ancillary files to execute, such as HTML documents that depend on image files or CSS stylesheets to render as intended, were evaluated on whether they opened as plain text.
- Programming or scripting files authored in plain text were tested as text files; they were not tested to see if the code they contained executed properly.
- Certain files deemed "unreadable" out of context of the associated files in their directory were considered to pass if they opened; for example, single-frame AVI files isolated from sequence.
- Package files, such as ZIP, passed if the package opened. The package contents were not tested.

4.3 Testing Timeline

The authors conducted testing over a five-month period from October 12, 2015 to March 23, 2016. The second author had a 13 hour per week appointment to the project, and conducted 97% of all initial tests. Prior to finalizing results, the primary author verified all files identified with status "fail" with the exception of those in the JPEG 2000 format (explanation to follow). During testing, ingest into the Medusa repository continued uninterrupted. The final population size reflects the number of files in Medusa on the final day of testing.

5. RESULTS

5.1 Overview

The authors tested 5,043 randomly sampled files[7] from a population of 8,209,807 (the population constituted the totality of files then housed in the Medusa repository). Statistically, this ensures to within a 2% margin of error and a 99% confidence level that the results are representative of repository-wide file format risk. Results, however, are not valid to within the margin of error for subpopulations of specific file formats. For example, the repository houses approximately 1.9 million files in the JPEG format (about 23% of all files), and indeed, approximately 1,141 files (about 22% of the sample set) were tested against the JPEG testing profile, ensuring a 4% margin of error for JPEG results at the desired 99% confidence level. On the other hand, the repository houses about 13,500 audio files with the format WAV (0.16% of all files), and tested 9 (0.18% of sample), meaning that the results are only valid to within a

[6] During testing, Medusa ran FITS version 0.8.3. FITS itself runs several metadata extractors such as Jhove (http://jhove.sourceforge.net) and DROID (http://www.dcc.ac.uk/resources/external/droid). FITS fields that accompany the full test data set are too numerous to list, but include PRONOM value, MIME type, file format name, file size, and last-modified-date.

[7] This paper presents snippet tabular views of project data; a comprehensive data-set is available at http://hdl.handle.net/2142/89994.

43% margin of error for the repository’s WAV files. While a future phase of research will focus on intensive testing within data strata such as file formats of interest, the authors acknowledge the limitations inherent to a purely random sample in this paper’s results.

As shown in Table 2 below, approximately 11% of files tested received a Fail status. While alarming at first glance, files failed to open for a variety of reasons, which are expanded on below.

Table 2. Results of testing by pass or fail

Status	Number	% of sample
Pass	4,479	89%
Fail	564	11%
TOTAL SAMPLE	5,043	(100%)

5.2 Triaging Results by File Format Profile

There isn’t a simple, programmatic way to triage test results by file format. One could sort by MIME type, PRONOM identity, or file format name, but these all represent different things. In the sample, FITS results show 47 MIME types, 67 PRONOM file formats (FITS reported no PRONOM value for 382 files, or about 8% of the sample set), and 77 file formats. However, the authors tested files against 93 Testing Profiles (see above), each one generally named after a file extension, and present these as the most consistent value for sorting data.

Table 3. Pass/Fail status for ten most frequently occurring file formats in sample

Testing Profile	Pass	Fail	Total Tested
TIFF	1276	1	1277
JPEG	1124	13	1137
JPEG2000	325	434	759
XML	540	2	542
PDF	402	0	402
GIF	192	3	192
HTML	130	0	130
TXT	114	0	114
EMLX	81	0	81
DOC	37	2	39

6. ANALYSIS

6.1 Files with Status Pass

Among files that passed muster, TIFF, PDF, and TXT performed especially well. 1276 out of 1277 TIFFs tested passed, as did all 402 PDFs and all 114 TXT files.

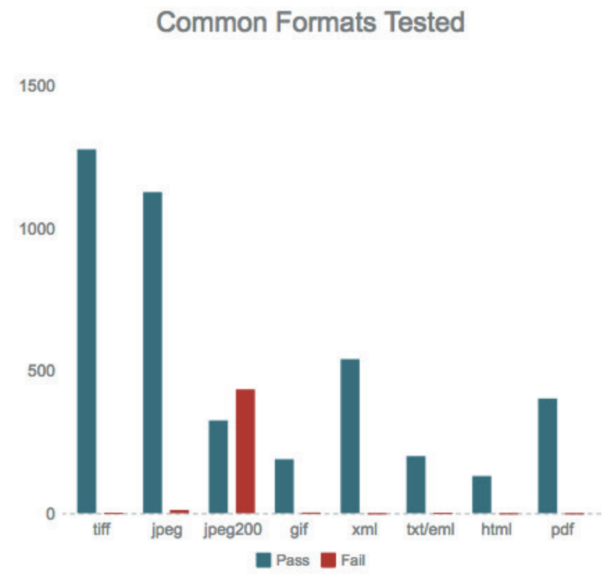


Figure 2. Pass/Fail for frequently occurring file formats in sample (visual representation based on Table 3)

6.2 JPEG 2000 Files with Status Fail

The majority of failed tests (434 of 564, or 77% of all tests with status Fail) occurred for files in the JPEG 2000 format, the third-most common file format in the repository behind TIFF and JPEG. To understand what this failure rate represents, some background on JPEG 2000 at Illinois is necessary. In 2007, the library adopted JPEG 2000 as its file format of choice for high-resolution preservation master image files produced in monographic digitization efforts, primarily to benefit from storage gains that JPEG 2000 lossless compression promised over the uncompressed TIFF alternative. The potential for JPEG 2000 to become a trusted format for access and preservation image files had at that point garnered considerable traction in the library field [7], and Illinois’ then-preservation managers felt confident enough to prefer JPEG 2000 to TIFF.

Acting on this policy, Illinois contracted with an off-site vendor to both deliver page image files of digitized items in the JPEG 2000 format, and to create a set of scripts to support the output of JPEG 2000 files in locally managed digitization workflows. As a result, Illinois took custody of hundreds of thousands of page images produced externally and in-house from 2007-2014, all using a related set of scripts to generate JPEG 2000 files.

While these image files are viewable in certain software applications, they are considered corrupt by others. FITS data on 100% of failed JPEG 2000 files confirms them as well-formed and valid to the format standard, a status bolstered by informal spot checks of several files using the JPLYZER[8] tool. In addition, the problematic JPEG 2000 files are able to render in certain open-source image manipulation software applications like ImageMagick[9] and Kakadu[10]. However, many consumer-grade software applications cannot open them, with Photoshop in particular throwing the error: “Could not complete your request because the file format module cannot parse the file.”

Experts in digital preservation have expressed concern that the nature of the JPEG 2000 standard would lead to this sort of problem. In 2011, van der Knijff wrote, “the current JP2 format specification leaves room for multiple interpretations when it comes to the support of ICC profiles, and the handling of grid resolution information. This has lead [sic] to a situation where different software vendors are implementing these features in different ways” [16]. While Illinois has not determined with certainty what variable differentiates its problematic JPEG 2000 files from those that open in Photoshop and other common software applications, it now knows that its repository houses hundreds of thousands of files that are unwieldy to many staff and patrons. The open source tools that can open these files without error are utilized primarily by specialists in file manipulation. They are not regularly employed by the library’s back-end users in its digitization lab or special collections units, nor by the scholars or graphic designers who frequently request image files from collection curators. When these users encounter such files, they most often find they cannot use them.

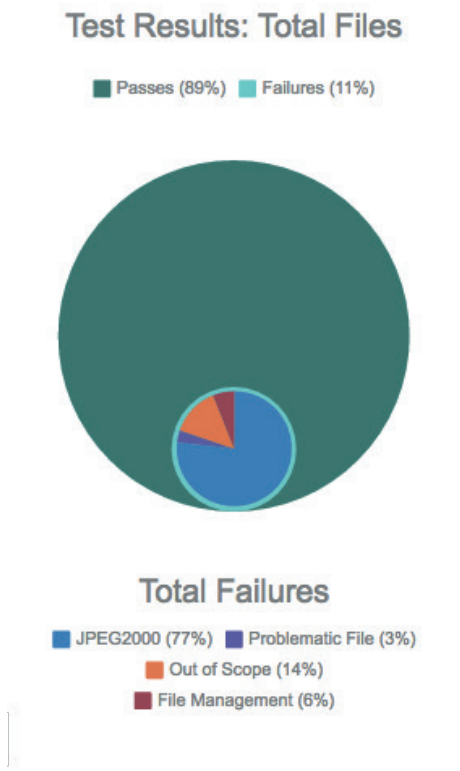


Figure 3. Percentage of Pass/Fail Test Status with Breakdown by Failure Type

6.3 Non-JPEG 2000 Failures

Files deemed to have failed to open according to their assigned profile did so for a variety of reasons, not all of which indicate file-format-based risk. In fact, by classing reasons for failure into groups *Out of Scope* (indicating they are not within the parameters of the testing regime), *Problematic File* (indicating the bit-stream itself is not readily openable), and *File Management* (indicating issues related to practices of naming and organizing files prior to their acquisition), the analysis below shows that only a small portion of non-JPEG 2000 failures are symptomatic of file format endangerment as it is generally understood.

6.3.1 *Overview of Non-JPEG 2000 Failures for Reason Out of Scope*

78 of the 130 non-JPEG 2000 files flagged as failures represent varieties of bit-streams that, while unfit to be opened and evaluated as discrete entities, are nonetheless currently retained by the repository as essential to their collections. 48 of them fell into the category of *System file not within scope of current testing*. Formats with this result included APMASTER, AUX, BAK, BIN, COM, DAT, DB, DLL, DS_STORE, EMLXPART, FRF, FRM, FRX, ICM, LOCK, MYD, PFB, PLIST, SCR, SYS, and V. These are predominately system files, executable files, and auxiliary files such as those created by software during data compilation, and belong overwhelmingly to born digital electronic records acquired by the University Archives. Most system and auxiliary files in these formats are not meant to be opened by a human computer user. (Executable files, on the other hand, frequently represent items of interest to patrons, and shall provide the focus of a future phase of research).

12 files fell into the category *Auxiliary file created and used by a software program, not meant to be opened as individual file*. Most of the files with this result were in the FRDAT format produced by AbbyFineReader software. FRDAT is a proprietary file format used by AbbyFineReader in digital imaging and optical character recognition workflows at Illinois. The files have been retained with a significant number of digitized book packages, although their long-term utility merits question.

11 files were temporary files with underscores, tildas, or dollar signs in their names that are not meant to be opened. Many repositories delete such files on ingest, but Medusa administrators have at present not adopted this practice for deposits. Specifically, 9 files fell in the category *Not meant to be opened--Mac system file with underscore in name,* 1 file fell in the category *Not meant to be opened - temporary file with ~$ in name*, and 1 file fell in the category *Not meant to be opened--software system file with @ symbol in name*.

Similarly, 5 bitstreams fell into the category *Not a file - artifact of disk formatting***.** These bitstreams registered with Medusa as files, although with names like FAT1 and FAT2 and sizes of 1KB, they are clearly artifacts of formatting on storage devices accessioned in collections of born digital electronic records.

Finally, 2 files failed testing with the reason, *Software available on market, but testers have not yet acquired it.* One was in the SAV format containing binary statistical data for the SPSS[11] platform. The other was a TBK file, a proprietary electronic learning platform file for software called ToolBook[12]. While the software to open these files exists for purchase on the market, in neither case did the testers procure it in time for publication.

6.3.2 *Overview of Non-JPEG 2000 Failures for Reasons Related to File Management Practices*

16 files fell in the category *No file extension*. Most of these were plain text files, frequently notes or works in progress, from collections of born digital personal records. Along similar lines, 2 files were appended with ad hoc file extensions and were given the failure reason *Not a file extension.* On closer inspection, these also turned out to be personal notes in collections of electronic records, where the depositor made up a file extension as a mnemonic device (e.g., authoring a text

[8] JPLYZER (http://jpylyzer.openpreservation.org/) is a “validator and feature extractor for JP2 images” produced by the EU FP7 project SCAPE (SCalable Preservation Environments).

[9] http://www.imagemagick.org/script/index.php

[10] http://kakadusoftware.com/

[11] http://www.ibm.com/analytics/us/en/technology/spss/

[12] http://www.sumtotalsystems.com/enterprise/learning-management-system/

document about a colleague and giving it an extension with that person's initials). While these files do not indicate file format endangerment, they do pose certain challenges to curation.

2 files were *Saved with incorrect extension*, both for unknown reasons. One was a JPEG with extension 000, and the other was a Microsoft Word file with extension 2_98, both of which files opened without a problem when appended with the correct extension. Both file formats were identified correctly by FITS.

More problematic are the 14 files that failed for the reason, *Despite file extension, file is in a folder designating it for another system purpose.* File formats with this result included GIF and JPEG—ostensibly image formats, although the files in question do not render as such, because they were created by a content management system for other purposes. Namely, numerous files from collections of born digital records acquired by the University Archives from former users of the FrontPage website authoring and management software contain files nested in a folder named "_vti_cnf". These software-generated folders contain files with the same names and extensions as JPEG and GIF files one level up in the directory hierarchy, but they are not in fact image files—rather, they were generated by FrontPage to keep track of versioning information of those files. Similarly, a JPEG file nested in folders called ".AppleDouble" indicate it to be a version tracking file used by an early Unix-like iteration of the Macintosh operating system. This "JPEG" does not render as an image file.

6.3.3 *Overview of Failures for Reason Problematic File*

18 non-JPEG 2000 files failed for reasons related to problematic file formatting.

13 failed for the reason, *Software considers file invalid*. 2 were JPEGs from the same collection of born digital electronic records, both with a last-modified-date in the year 2000. In attempting to open them, Photoshop provided the error: "Could not complete your request because a SOFn, DQT, or DHT JPEG marker is missing before a JPEG SOS marker." These files were generated by a little-known (though apparently still available) software called CompuPic(R)[13]. The other 11 files in this category have the WMZ extension, and appear to be compressed images from a slide presentation (the Windows operating system thinks they are Windows Media Player skin files, but some web research[14] shows that Microsoft Office software has used the WMZ extension for other purposes in the past; at present, testers have had no success opening WMZ files in the Medusa repository). The WMZ files in question were created in 2001, and also belong to a collection of born digital electronic records.

3 files failed for the reason, *File does not render in software*. Two are document files, one in the Microsoft Word DOC format, and the other in RTF. Embedded technical metadata in both files suggests they were created, at an indeterminate date, by an instance of Corel WordPerfect. Both files originate from a collection of born digital electronic records. The third file in this category is a GIF from a collection of born digital electronic records that appears to have been corrupt at the time of deposit, as it is in a folder of GIF files, and the others open without fail.

Table 4. Number of Test Failures by Reason and Type of Reason for all non-JPEG 2000 Failures

Reasons	Reason Type	Total
System file not within scope of current testing	out of scope	48
Auxiliary file created and used by a software program, not meant to be opened as individual file	out of scope	12
Not meant to be opened—Mac system file with underscore in name	out of scope	9
Not a file—artifact of disk formatting	out of scope	5
Software available on market, but testers have not yet acquired it	out of scope	2
Not meant to be opened—software system file with @ symbol in name	out of scope	1
Not meant to be opened - temporary file with ~$ in name	out of scope	1
TOTAL OUT OF SCOPE		***78***
No file extension	file management	16
Despite file extension, file is in a folder designating it for another system purpose	file management	14
Not a file extension	file management	2
Saved with incorrect extension	file management	2
TOTAL FILE MANAGEMENT		***34***
Software considers file invalid	problematic file	13
File does not render in software	problematic file	3
Software unavailable	problematic file	1
Software attempts to convert file to new version of format and fails.	problematic file	1
TOTAL PROBLEMATIC FILE		***18***
TOTAL ALL CATEGORIES		***130***

1 file failed for the reason, *Software unavailable*. This was in the format 411, a proprietary thumbnail image format for early Sony digital cameras, and originated from a collection of born digital electronic records.

1 file failed for the reason, *Software attempts to convert file to new version of format and fails*. This is a Corel WordPerfect WPD file that cannot be opened in the latest version of WordPerfect. It originated from a collection of born digital electronic records.

7. DISCUSSION

Success and failure rates reflected in this study's results do not necessarily bespeak the preservation viability of specific file formats over others. Frequently they reflect the practices of the community of users who produced them, or the circumstances under which they were created. For example, problematic files in the sample were often either produced using software that never established a broad user base, or were output by one company's software but in a competitor's proprietary format (e.g. unreliable RTF and DOC files created by WordPerfect). In the case of perennially reliable file formats like TIFF, PDF, and TXT, however, a strong support system has emerged around them, with consistent software support across multiple operating systems.

[13] A trial version is still available for download at http://www.photodex.com/compupic, but the software was created in 2003 and does not successfully install in the Windows 7 environment.

[14] http://stackoverflow.com/questions/3523083/decompress-wmz-file

7.1 JPEG 2000 Policy

In contrast to its TIFF holdings, the repository houses a number of JPEG2000 files (approximately 700,000, to extrapolate from the failure rate into the entire subpopulation of files with extension JP2) whose image bit-streams are intact, but whose file structure makes them inaccessible in common image management software. These files do not pose an immediate preservation risk, as it is well within the institution's ability to reformat them without loss [10]; rather, they pose a genuine access hurdle for many users.

Due to frustration with managing files in the JPEG 2000 file format as reflected in this research, the Illinois library has shifted its practices around the stewardship of preservation master files back to TIFF. The library, however, has not abandoned the JPEG 2000 format entirely—rather, it is limiting the scope of its use. Despite its drawbacks, JPEG 2000 has distinguished itself as particularly advantageous for online image presentation systems, thanks to the speed and efficiency with which web applications retrieve and render high-resolution JPEG 2000 images. In digital libraries, JPEG 2000 has found its home in the back-end of many image presentation systems, particularly those that serve millions of pages of library content online (both Chronicling America[15] and the HathiTrust Digital Library[16] rely on JPEG 2000 for serving page images). Likewise, the Illinois library is using JPEG 2000 as a back-end presentation format in its own locally managed digital image collections[17], while retaining preservation master files for digital images in the TIFF format.

7.2 Born Digital Electronic Records

Electronic records make up only a small slice of Medusa's collections (about 2 TB out of 60), but their files are disproportionately represented in failed tests. The 52 non-JPEG 2000 files that failed testing for reasons of questionable *File Management* practices (34) and for the reason *Problematic File* (18) constitute 1% of the sample set, and originate overwhelmingly from collections of born digital electronic records. This suggests that the curation of born digital collections represents a hot spot, so to speak, warranting the attention of local preservation managers.

Collections of born digital electronic records acquired by the University Archives and collections of digitized collections from departmental libraries, however, often have different curatorial needs. In the sample, the authors discovered the 411 format used by an early Sony digital camera called the Mavica[18]. Because proprietary rendering software for 411 files is presently unavailable without going to great lengths, the tested 411 file (created in 2002) was given a "Fail" status as unopenable. Some would say that such a file ought to be discarded on ingest and not retained at all—after all, if usable thumbnails are needed, they can be generated from the full-size image files stored in the same folder. However, the model name "Mavica" does not show up in any of the technical metadata for the full-size JPEG from which this thumbnail was derived, and the only way to know that this camera was used at all is *because* the associated thumbnail file with extension 411 was retained in the repository. From this perspective, the 411 file possesses potential research value. It provides evidence of the camera the person who took the photo used. It also demonstrates how an early digital camera platform generated thumbnail images. A technically useless file, it nevertheless provides historical context to the creation of other files in the collection, ensuring an unbroken "archival bond[19]" between bit-streams.

This suggests a need for different retention policies for different types of content within the repository. While curators of digitized monographs may look approvingly on disposing of "noise"—wiping the slate clean of artefacts of former image display software, system-generated files, and the like—an archivist may prefer a more conservative file retention policy for collections of born digital records, since these files may well provide insight into the creation and use of other files, or even help a researcher judge the authenticity of files as records.

7.3 Limitations of Methodology

The random sampling method, as employed by this study, poses certain limitations on the relevance of results to specific subpopulations of data, and implies the need for future work. The Medusa repository's collections originated from a variety of sources and workflows, some of which have produced more files than others. This means that image formats from book digitization efforts occurred much more frequently in the sample than audio formats from the library's nascent media preservation program, and that files from vendor-digitized general collections appeared with greater frequency than those from born digital special collections. By analyzing a random sample of files across a repository of highly disparate subpopulations of data, results provide an initial assessment of risk that is only statistically meaningful from a bird's eye view.

More importantly, the authors find the testing methodology described in this paper to be useful only as a blunt instrument for assessing barriers to content access. While other institutions may find a similar exercise useful, it is the authors' hope that their experiment will serve as a preliminary step toward elaborating a more sophisticated and effective means of assessment.

8. NEXT STEPS

Based on this study, the authors recommend that Medusa's digital preservation managers 1) isolate problematic JPEG 2000 files, particularly those that demonstrate high use, and remediate to TIFF format, and 2) devise an improved methodology for a follow-up study focused exclusively on collections of born digital electronic records, with an eye toward appraisal policy development and enhanced repository services for them.

9. CONCLUSION

The testing and analysis process detailed in this paper has forced Illinois preservation managers to identify and confront

[15] http://chroniclingamerica.loc.gov/

[16] HathiTrust (http://hathitrust.org/) relies on JPEG 2000 as a preservation format as well, but notably enforces formatting requirements on ingest, ensuring strict technical uniformity of all image files [10].

[17] Currently in beta at https://digital.library.illinois.edu/, the digital library utilizes the IIIF (http://iiif.io/) image interoperability framework, which allows for on-the-fly conversion and delivery of access derivatives in a variety of image formats to patrons, largely obviating the problem of keeping a single "master" file on hand in a readily accessible format to deliver to patrons in need.

[18] http://fileformats.archiveteam.org/wiki/Sony_Mavica_411

[19] http://www2.archivists.org/glossary/terms/a/archival-bond#.V4P-4vkrJmM

genuine problems curators and patrons face when attempting to open and use files stewarded in the Medusa repository. In the absence of similar studies, it is difficult to know whether Illinois' specific challenges are generalizable to those experienced by other institutions. Nevertheless, the testing method and findings presented here ought to prove useful to other researchers and managers interested in taking an evidence-based approach to assessing barriers to file rendering in digital preservation repositories.

10. ACKNOWLEDGMENTS

The authors would like to thank the University of Illinois at Urbana-Champaign Campus Research Board for generously supporting this project. In addition, they'd like to express their gratitude to Howard Ding for providing invaluable technical expertise and code development critical to its success.

11. REFERENCES

[1] Anderson, R., Frost, H., Hoebelheinrich, N., & Johnson, K. (2005). The AIHT at Stanford University: Automated Preservation Assessment of Heterogeneous Digital Collections. *D-Lib Magazine*, *11*(12), 10. http://doi.org/10.1045/december2005-johnson

[2] Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., & Hofman, H. (2009). Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries*, *10*(4), 133–157. http://doi.org/10.1007/s00799-009-0057-1

[3] Cochrane, E. (2012). *Rendering Matters - Report on the results of research into digital object rendering*. Archives New Zealand. Retrieved from http://archives.govt.nz/rendering-matters-report-results-research-digital-object-rendering

[4] De Vorsey, K., & McKinney, P. (2010). Digital Preservation in Capable Hands: Taking Control of Risk Assessment at the National Library of New Zealand. *Information Standards Quarterly*, *22* (2), 41–44.

[5] Holden, M. (2012). Preserving the Web Archive for Future Generations. In *The Memory of the World in the Digital age: Digitization and Preservation* (pp. 783–795). Vancouver: United Nations Educational, Scientific, and Cultural Organization. Retrieved from http://ciscra.org/docs/UNESCO_MOW2012_Proceedings_FINAL_ENG_Compressed.pdf

[6] Illinois Digital Environment for Access to Learning and Scholarship. (n.d.). FormatRecommendations. Retrieved July 30, 2013, from https://services.ideals.illinois.edu/wiki/bin/view/IDEALS/FormatRecommendations

[7] Kulovits, H., Rauber, A., Kugler, A., Brantl, M., Beinert, T., & Schoger, A. (2009). From TIFF to JPEG 2000?: Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings. *D-Lib Magazine*, *15*(11/12). http://doi.org/10.1045/november2009-kulovits

[8] Pennock, M., Wheatley, P., & May, P. (2014). Sustainability Assessments at the British Library: Formats, Frameworks, and Findings. In *iPres2014: Proceedings onf the 11th International Conference on Preservation of Digital Objects* (pp. 142–148).

[9] Rieger, O. Y. (2008). *Preservation in the Age of Large-Scale Digitization: A White Paper*. Washington, D.C.: Council on Library and Information Resources. Retrieved from http://www.bib.ub.edu/fileadmin/fdocs/pub141.pdf

[10] Rimkus, K., & Hess, K. (2014). HathiTrust Ingest of Locally Managed Content: A Case Study from the University of Illinois at Urbana-Champaign. *The Code4Lib Journal*, (25). Retrieved from http://journal.code4lib.org/articles/9703

[11] Rimkus, K., Padilla, T., Popp, T., & Martin, G. (2014). Digital Preservation File Format Policies of ARL Member Libraries: An Analysis. *D-Lib Magazine*, *20* (3/4). http://doi.org/10.1045/march2014-rimkus

[12] Rog, J., & Van Wijk, C. (2008). Evaluating file formats for longterm preservation. *Koninklijke Bibliotheek*, *2*, 12–14.

[13] Rosenthal, D.S.H. (2010). Format obsolescence: assessing the threat and the defensesnull. *Library Hi Tech*, *28*(2), 195–210. http://doi.org/10.1108/07378831011047613

[14] Ryan, H. (2014). Occam's Razor and File Format Endangerment Factors. In iPres2014: Proceedings of the 11th International Conference on Preservation of Digital Objects (pp. 179–188).

[15] Stanescu, A. (2005). Assessing the durability of formats in a digital preservation environment: The INFORM methodology. *OCLC Systems & Services: International Digital Library Perspectives*, *21*(1), 61–81. http://doi.org/10.1108/10650750510578163

[16] van der Knijff, J. (2011). JPEG 2000 for Long-term Preservation: JP2 as a Preservation Format. *D-Lib Magazine*, *17*(5/6). http://doi.org/10.1045/may2011-vanderknijff

[17] Waters, D., & Garrett, J. (1996). *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*. The Commission on Preservation and Access and The Research Libraries Group. Retrieved from http://www.clir.org/pubs/reports/pub63watersgarrett.pdf

Practical Analysis of TIFF File Size Reductions Achievable Through Compression

Peter May
British Library
96 Euston Road
London
Peter.May@bl.uk

Kevin Davies
British Library
Boston Spa
West Yorkshire
Kevin.Davies@bl.uk

ABSTRACT

This paper presents results of a practical analysis into the effects of three main lossless TIFF compression algorithms – LZW, ZIP and Group 4 – on the storage requirements for a small set of digitized materials. In particular we are interested in understanding which algorithm achieves a greater reduction in overall storage, and whether there is any variation based on the type of file (e.g. colour depth). We compress 503 files with two software utilities – ImageMagick and LibTiff – and record the resulting file size for comparison against the original uncompressed version. Overall we find that in order to effectively (although not necessarily optimally) reduce total storage, Group 4 compression is most appropriate for 1-bit/pixel images, and ZIP compression is suited to all others. We also find that ImageMagick – which uses the LibTiff library – typically out-performs LibTiff with respect to compressed file sizes, noting that this appears to be the result of setting the "Predictor" tag.

Keywords

TIFF; Compression; LZW; ZIP; Group 4; ImageMagick; LibTiff

1. INTRODUCTION

Tagged Image File Format (TIFF) is considered the de facto format for preservation master image files, providing a simple tagged structure for storing raster image pixel data along with associated metadata. The format itself is stable and well documented, with the specification having not seen a major revision since 1992 [9]. It is also widely adopted, both in terms of graphics software and in terms of Library and Archive adoption. The British Library is no exception to this, having received around 5 million TIFF files through our Endangered Archives Programme alone.

TIFF files can be very large, however, leading to storage cost problems for big collections, and potentially impacting on the long-term preservation of these and other collections for financial reasons; the larger the files, the fewer that can be stored within a defined storage system.

One approach to mitigate this, whilst retaining use of the TIFF format, would be to compress the image payload data. TIFF enables this by supporting a variety of different compression algorithms, such as the lossless Group 4, LZW and ZIP algorithms. Being lossless, these algorithms all enable a TIFF image to be reduced in size, with the image data being fully retrievable at a later stage (through decompression).

From a storage perspective though, it is not clear what impact each of these compression approaches has on the overall size of a stored TIFF collection, particularly for the types of digitized files held by libraries and other memory institutions. Does one algorithm compress the files to a greater extent than another? Are different algorithms suited to different types of file?

In addition to this, compression is applied through the use of a particular software application/library, such as ImageMagick[1] or LibTiff[2]. Does the choice of software impact on the amount of compression achievable?

This paper reports on a practical experiment performed at the British Library analyzing the effects of LZW and TIFF compression on the storage size of a small set (503 files) of digitised material. It focuses on the average file sizes achievable through these compression algorithms, across different image colour depths, and through using two popular and freely available software utilities for performing the compression (the previously mentioned LibTiff and ImageMagick).

We start by briefly outlining the background to TIFF files, their overall structure and details about community recommendations on the use of TIFF files, particularly with respect to compression. Section 3 then describe the experimental methodology applied, covering details about the process, hardware and software, and the dataset used. The results are presented and analysed in Section 4, with discussion about what this means in practice outlined in Section 5.

2. TIFF FILES AND THEIR USE

TIFF is a bitmap image format originally created by the Aldus Corporation in the mid-1980's, but now owned by Adobe after they acquired Aldus in 1994 [9]. It evolved from a bitonal format to encompass grayscale and full-colour image data, as well as support for a variety of compression technologies.

The current specification (revision 6) [9] is split into two parts; part 1 describes baseline TIFF, which covers the core parts essential for TIFF readers to support. Part 2 covers extensions to the baseline, covering features which may not be supported by all TIFF readers.

2.1 Structure and Compression

TIFFs are tag-based in structure. They start with an 8 byte header which contains an offset value to the first Image File Directory (IFD) containing tags (such as height, width, image data location, etc.) and associated values pertaining to the first image within the file. An offset tag to the next IFD allows another sub-image to be included (e.g. the next page, or a thumbnail) in the same manner, and so on. Baseline TIFF readers are not required to read beyond the first IFD however.

In addition to providing the location of the image data within the file, tags also provide details about the compression applied to that data. It should be noted that, as stated in the TIFF specification, "Data compression applies only to raster image data. All other TIFF fields are unaffected" [9].

Baseline rev. 6 TIFF images can be compressed using either the lossless Modified Huffman Compression algorithm for bi-level images, or the lossless PackBits compression algorithm (both are described in the specification). Extensions to the baseline

[1] http://www.imagemagick.org/script/index.php

[2] http://www.libtiff.org/

TIFF define additional compression schemes though: Group 3 and Group 4 for bitonal images (both lossless), as well as LZW (Lempel-Ziv & Welch; lossless) and JPEG (lossy). Compression enhancements for ZIP (Deflate/Inflate; lossless) and 'new style' JPEG were specified in supplementary TIFF Technical Notes [1].

LZW was originally defined as a baseline compression scheme in TIFF version 5, but was moved to the extensions section in TIFF version 6 due to licensing issues surrounding LZW patents. These patents expired in 2003/2004 (US/Europe respectively) [3] effectively removing the need for legal-related restrictions on the use of LZW compression [10].

2.2 Community Guidelines on use of TIFFs

TIFF files are widely used in libraries and archives as master files for digitized still images. Recommendations for their use for this purpose are quite consistent, typically recommending uncompressed or LZW compressed images.

The Succeed Project assessed existing digitization recommendations, providing a summary of these and consolidating them into their own recommendations [5]. TIFF v6 was the recommended master file format for still images, either uncompressed or using LZW compression.

The U.S. National Archives and Records Administration (NARA) Technical Guidelines for Digitizing Archival Materials for Electronics Access suggest LZW or ZIP lossless compression could possibly be used in an actively managed digital repository. JPEG compression should not be used [5].

The same LZW or ZIP compression recommendation is also true for the Federal Agencies Digitization Guidelines Initiative (FADGI) 2010 guidelines for digitizing cultural heritage materials (although uncompressed is preferred) [7]. This is unsurprising given they essentially derive from the NARA guidelines.

Other guidelines are more restrictive on the use of compression, effectively prohibiting it. For example, the National Digital Newspaper Program (NDNP) guidelines state that master page images should be delivered as uncompressed TIFF v6.0 files, and supported by derivative JPEG2000 files for end user access [8].

The British Library's internal digitization guidelines are also consistent with those from the wider community, recommending no compression or LZW compression for TIFF (v6) files.

These guidelines appear to be trying to balance long-term preservation accessibility (though minimizing complications by using no compression) with reduced storage (through lossless compression). In terms of storage reduction however, it is not always clear from the recommendations why a particular algorithm is chosen. More so, if the aim of recommending compression is to reduce storage requirements, is the algorithm choice sufficient?

Gillesse *et al*, at The National Library of the Netherlands (KB) undertook a research project looking at potential alternatives to TIFF Master Image files, comparing LZW compressed TIFF with JPEG2000, PNG and JPEG [2]. They found that based on their two test sets of ~100 originals, "it appears that TIFF LZW in lossless mode can yield a benefit of about 30% compared to an uncompressed file" [2]. This is a useful indication of the amount of storage that can be saved but, being derived from a small test sample, how accurate is it? And what variation, if any, is there based on the type of content tested?

Evidence is not easy to find, and is often embroiled in other investigations and disciplines, particularly medical related [4].

Anecdotal evidence available on the internet[3] suggests that we should expect variation in the compressibility of files based on the amount of detail within the image and the colour depth. However such reports typically only test a handful of files, and provide limited – if any – detail of the methodology taken; hardly conclusive evidence.

3. METHODOLOGY

Figure 1 depicts the overall process used to compress the set of files described below using LZW and ZIP algorithms, interrogate the files to obtain relevant image properties, and compile the results into a CSV file suitable for detailed analysis. This process was automated through a shell script.

3.1 Data Set

503 TIFF images were randomly taken from our Endangered Archive Programme's submissions. These comprised a variety of bit-depth images as detailed in Table 1, and covered a broad range of categories such as books, magazines, microfilm, newspapers and photographs.

Table 1: Sample set details grouped by bit-depth

Bit Depth	File Count	Group 4 Compressed	Total Size
1	56	Yes	5.5 MiB
8	57	-	1231.1 MiB
24	345	-	8512.3 MiB
48	45	-	1636.4 MiB
Total:	***503***		***11385.3 MiB***

3.2 Data Preparation

As can be seen, the sample of TIFF files used were largely uncompressed; the only compressed files were a selection of 1-bit/pixel microfilm records, compressed using the Group 4 algorithm. These files were first decompressed using the 'tiffcp' utility before the main conversion was performed.

3.3 Compression Software

Uncompressed TIFFs are compressed (and subsequently decompressed), as shown in Figure 1, using either the ImageMagick or LibTiff versions mentioned below. These software utilities were chosen as they are commonly used for image file manipulation, particularly on Linux environments. In both cases, standard installations and default settings are used.

Other versions of these utilities, and other graphics software such as Photoshop, have not been investigated.

ImageMagick (6.6.9.7-5ubuntu3.4):

Used to compress and decompress files using ZIP and LZW algorithms. It was also used to obtain image properties such as bit depth, dimensions and number of unique colours.

- convert -compress zip "<inputfile>" "<outputfile>"
- convert -compress lzw "<inputfile>" "<outputfile>"
- convert -compress none "<inputfile>" "<outputfile>"

Note: ImageMagick depends on LibTiff (using the same version as below, in our case) for TIFF manipulation. As we will see, results still vary between standalone LibTiff and ImageMagick.

[3] http://havecamerawilltravel.com/photographer/tiff-image-compression

http://www.scantips.com/basics9t.html

http://videopreservation.conservation-us.org/tjv/index.html

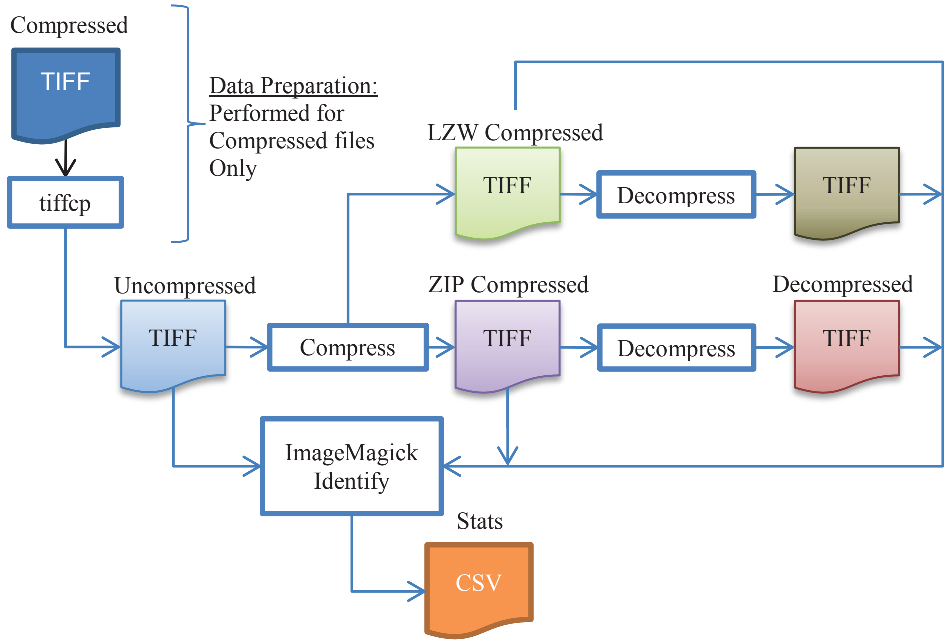


Figure 1: The process used to compare file sizes between the uncompressed, compressed and decompressed TIFFs.

LibTiff (libtiff4-3.9.5-2ubuntu1.9):

LibTiff's 'tiffcp' utility was also used to compress/decompress files, and to remove existing Group 4 compression from files.

- tiffcp -c zip "<inputfile>" "<outputfile>"
- tiffcp -c lzw "<inputfile>" "<outputfile>"
- tiffcp -c none "<inputfile>" "<outputfile>"

Note: Group 4 compressed images were taken as is from the original sample, and not recompressed.

3.4 Hardware

The compression and analysis process was executed on an Ubuntu 12.04.2 64bit (Kernel: 3.5.0-54-generic x86_64) VM running on a HP ProLiant DL385p Gen8 server. The VM was allocated 1 CPU (2.294GHz), ~6GB of RAM and ~500GB storage.

3.5 Entropy Calculations

As part of our analysis we calculated the average image entropy as a measure of how "busy" an image is. To do this we used Fred Weinhaus'[4] 'entropy' script, which uses ImageMagick to extract the histogram for each colour channel of an image, and determine the distribution of each colour value within that histogram. Normalisation of these distributions gives an entropy value of between 0 (a solid plane of colour) and 1 (a uniform gradient of all possible colour) for each channel. The average of the entropy values for all colour channels is used as the average entropy for the image.

3.6 Uncompressed vs. Decompressed Pixels

Pixel data from the original uncompressed TIFF files and the decompressed files were compared using ImageMagick's 'compare' command with '–metric ae' option, which measures the number of pixels differing between the two images. In all cases, pixel data in the original and decompressed TIFFs was identical.

3.7 Reported File Sizes

Compression of a TIFF file is applied to the image payload only, however changes will often occur within other areas of the file (i.e. the tags) to describe this compression. Furthermore, software libraries applying the compression may affect, for example remove, other metadata within the file. File sizes reported in this paper are for the complete file, encompassing all changes made by the application of compression, as this best reflects the total storage requirements. From a preservation perspective however, all changes caused by compression should be considered.

4. RESULTS

This section presents the results from experimentation on the specified sample of collection material, with an analysis of the main findings.

The results are organized in a logical order following the process diagram shown in Figure 1 evaluating:

- Original Group 4 compressed files compared to their uncompressed "original" form.
- LZW/ZIP compressed files compared to their uncompressed "original" form.
- LZW/ZIP compressed files compared to their Group 4 compressed form (for 1-bit files).

4.1 Group 4 Compressed vs. Uncompressed File Sizes

Of the original sample of files, all 56 of the 1-bit TIFFs were found to be compressed with Group 4 compression. The initial step in our process decompressed these to present a uniform, uncompressed sample of files.

Table 2 shows the mean average ratio in file sizes between the Group 4 compressed files and their uncompressed counterparts. LibTiff's "tiffcp" utility was always used to decompress originally compressed TIFF files, and so no comparable results are available for the ImageMagick tool.

Table 2: Minimum, mean and maximum ratio of Group 4 file sizes with respect to their uncompressed size (to 1 d.p.)

Min	Mean	S.D.	Max
0.05%	12.68%	13.86%	47.15%

As can be seen, and as to be expected, the 1-bit Group 4 compressed TIFFs are smaller than their uncompressed counterparts, averaging ~13% of the uncompressed size. At most, the least compressed file is still over 50% smaller than its uncompressed form.

[4] http://www.fmwconcepts.com/imagemagick/entropy/index.php

Summary:

- Group 4 compressed files appear to be at least half the size of their uncompressed form.

4.2 LZW/ZIP Compressed vs. Uncompressed File Sizes

With all 56 Group 4 files de-compressed, the 503 uncompressed files become the base sample for further compression analysis. These are compressed using either LibTiff's 'tiffcp' command or ImageMagick's 'convert' command, and the file sizes recorded. Table 4 shows, for both software libraries, the minimum, maximum and mean average file sizes (in MiB[5]) for the original uncompressed files, and the resulting LZW or ZIP compressed files.

4.2.1 Effect of Compression Algorithm

With respect to compression algorithm, Table 4 shows three things. Firstly, irrespective of bit-depth and software utility, both ZIP and LZW compression generate compressed files with a mean average size smaller than the original uncompressed files.

This is also highlighted in Table 3, which indicates that LZW files are an average of ~51% or ~70% the size of the uncompressed original (for ImageMagick and LibTiff respectively), and ZIP files are an average of ~44% or 58% (respectively).

Ratios are calculated on a file-by-file basis across the entire 503 uncompressed sample files before averaging.

Table 3: Ratio of LZW/ZIP compressed file sizes as a percentage of the original uncompressed files (to 1 d.p.)

Library	Alg.	Min	Mean	S.D.	Max
ImageMagick	LZW	2.3%	51.2%	26.8%	133.8%
	ZIP	0.5%	43.9%	21.1%	99.4%
LibTiff	LZW	2.0%	69.8%	27.6%	130.0%
	ZIP	0.5%	58.2%	24.3%	98.5%

Secondly, generating smaller files from the use of compression is not guaranteed. A maximum compressed-to-uncompressed ratio being greater than 100%, as seen in Table 3, indicates that there are incidents where applying LZW compression using either software utility actually increases the file size. This predominantly affects 24-bit images in our sample, as summarised in Table 5, with 28 compressed images being larger than the original when using ImageMagick, compared to 68 when using LibTiff.

Table 4: Count of files, per bit depth and software library, whose LZW compressed size is greater than their original uncompressed size

Bit Depth	File Count	
	ImageMagick	LibTiff
8	-	1
24	28	68

Why should LZW compression increase the file size however? The original, and common, choice for LZW code table is to store 4096 entries, requiring 12-bits to encode every entry (2^12=4096). The initial 256 entries are reserved for the single byte values 0-255, while the remaining entries correspond to multi-byte sequences. Savings are made when sequences of bytes in the original file can be encoded by one 12-bit code; however, if this is not possible, then the 12-bit code for individual bytes is used instead, adding a 50% overhead to each byte. This is a simplified example, but illustrates the point of how LZW could create larger files.

Thirdly, these results highlight that, again irrespective of bit-depth and software utility, ZIP compression generates an average compressed file size smaller than that produced with LZW compression. This appears to be consistently true for our tested sample. Comparing the ratio of ZIP to LZW compressed file sizes on a file-by-file basis (shown in Table 6), ZIP compressed files are between ~22% and ~96% the size of LZW compressed files, with an average of ~84%. No individual ZIP file has therefore exceeded the size of the corresponding LZW file; if it had, the maximum ratio would have been larger than 100%.

Table 5: Minimum, maximum and mean average file sizes[5] for each colour depth grouping (to 1 d.p.; † to 1 s.f.).

		ImageMagick				LibTiff			
Bit Depth	TIFF*	Min (MiB)	Mean (MiB)	S.D. (MiB)	Max (MiB)	Min (MiB)	Mean (MiB)	S.D. (MiB)	Max (MiB)
1	Original	0.7	0.9	0.6	3.9	0.7	0.9	0.6	3.9
	LZW	0.03†	0.1	0.06†	0.3	0.02†	0.1	0.06†	0.3
	ZIP	0.006†	0.1	0.05†	0.2	0.006†	0.1	0.05†	0.2
8	Original	1.4	21.6	18.2	63.3	1.4	21.6	18.2	63.3
	LZW	0.4	13.3	15.5	55.3	0.5	17.0	17.8	62.6
	ZIP	0.4	11.4	12.7	45.1	0.5	15.2	15.2	52.8
24	Original	8.6	24.7	12.8	54.4	8.6	24.7	12.8	54.4
	LZW	2.4	16.4	14.2	60.5	4.5	21.2	14.0	57.1
	ZIP	2.0	13.9	11.1	45.0	3.4	17.5	11.6	43.3
48	Original	25.9	36.4	15.0	57.3	25.9	36.4	15.0	57.3
	LZW	6.8	9.5	2.6	14.9	12.7	16.8	5.0	25.8
	ZIP	5.6	7.6	2.0	11.9	9.9	12.8	3.1	19.2

* "Original" TIFF refers to the original uncompressed image.

[5] File size results are expressed in IEC 80000-13 binary prefixes; 1 KiB (kibibyte) = 1024 Bytes, 1MiB (mibibyte) = 1024 KiB.

Table 6: Minimum, mean and maximum ratio of ZIP to LZW compressed file sizes for each software library (to 1 d.p.)

Library	Min	Mean	S.D.	Max
ImageMagick	22.4%	86.0%	7.5%	95.6%
LibTiff	25.9%	82.6%	9.9%	95.5%
All	22.4%	84.3%	9.0%	95.6%

Summary with respect to Compression Algorithm:

- Either algorithm generates an *average* compressed file size smaller than the original, uncompressed average file size
- ZIP generates compressed files smaller than that produced with LZW.
- The LZW algorithm is capable of increasing the file size, rather than decreasing it.
- The ZIP algorithm has not, for this sample, increased the file size.

4.2.2 Effect of Bit-Depth

The ratios of compressed to original file sizes shown in Table 3 can be examined further based on the bit-depth of the original image. Results from this bit-depth analysis are shown below in Table 7.

These results are clearer at showing for which bit-depth LZW compressed files are not guaranteed to be smaller than their uncompressed originals (specifically, 8 and 24-bit).

They also reinforce, at each bit-depth level, the previously mentioned findings that the average ZIP compressed files are smaller than LZW compressed files. As per Table 6, Table 8 confirms this on a file-by-file basis, with ZIP compressed files being at most ~96% the size of the LZW compressed files.

Table 7: Ratio of LZW/ZIP compressed file sizes as a percentage of the original uncompressed file sizes for each bit-depth (to 1 d.p.)

	Alg.	Bit Depth	Min	Mean	S.D.	Max
ImageMagick	LZW	1	2.3%	17.1%	8.9%	34.2%
		8	24.6%	48.5%	19.0%	87.4%
		24	22.9%	60.3%	25.0%	133.8%
		48	17.5%	27.5%	4.4%	34.0%
	ZIP	1	0.5%	14.1%	7.5%	27.4%
		8	22.6%	42.4%	15.2%	71.2%
		24	19.0%	51.8%	18.2%	99.4%
		48	13.3%	22.2%	3.8%	27.7%
LibTiff	LZW	1	2.0%	16.7%	8.8%	33.8%
		8	32.6%	64.0%	22.6%	102.2%
		24	39.8%	82.2%	18.2%	130.0%
		48	35.0%	48.3%	5.7%	54.6%
	ZIP	1	0.5%	14.5%	7.7%	28.0%
		8	30.4%	57.9%	19.8%	89.2%
		24	22.2%	68.0%	18.0%	98.5%
		48	21.6%	37.9%	7.2%	46.6%

Table 7 shows that the average compression achieved varies with bit-depth, lessening as the bit-depth increases. For example, the average LZW compressed file size produced by ImageMagick is ~17% (of the uncompressed size) for 1-bit, ~48% for 8-bit, and ~60% for 24-bit. Interestingly though, 48-bit images appear to achieve substantially more compression than 8 and 24-bit images, with average compressed file sizes ranging between 22% and 48% of the original uncompressed size. Sample sizes should always be borne in mind, however if considered representative of a larger population value, then this indicates better compression performance on the larger sized image payloads afforded by the 48-bit colour depth.

Table 8: Minimum, mean and maximum ratio of ZIP to LZW compressed file sizes for each software library (to 1 d.p.)

	Bit Depth	Min	Mean	S.D.	Max
ImageMagick	1	22.4%	78.1%	15.5%	90.1%
	8	81.5%	88.4%	3.0%	92.2%
	24	74.1%	87.6%	4.9%	95.6%
	48	75.0%	80.4%	1.6%	82.7%
LibTiff	1	25.9%	82.9%	15.2%	95.1%
	8	84.2%	90.7%	3.1%	95.5%
	24	50.3%	81.9%	9.2%	91.8%
	48	61.6%	77.8%	6.6%	85.8%

Table 7 also clearly illustrates that 1-bit images are capable of being heavily compressed, more so than the other bit-depths. ZIP especially, is able to reduce these bitonal files to 0.5% of their original uncompressed size. Analysis of these 1-bit files shows that the heavily compressed ones have lower average image entropies (described in Section 3.5) than the less compressed files – see Figure 2.

Entropy is a quality indicating how "busy" an image is. Low entropy images – such as a photograph of the night sky, or a page of plain black text against a white background - contain large swathes of pixels with the same or similar colour values. Higher entropy images – such as a photograph of a crowd - have a great deal of contrast between neighbouring pixels.

The theory is that high entropy images have a greater variety of information (e.g., more variation in colour) to encode than lower entropy images, and therefore should be more difficult to compress effectively. These results support this theory for 1-bit images – the ability to Group 4 compress a 1-bit image appears to degrade as image entropy increases.

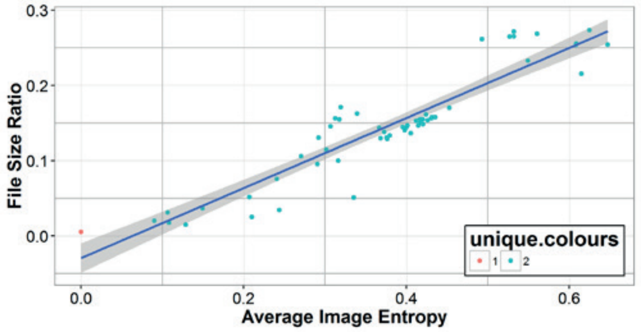


Figure 2: ImageMagick's ZIP to Uncompressed file size ratio for 1-bit TIFFs vs. the average image entropy

Finally, comparing ZIP compressed file sizes relative to LZW compressed sizes (rather than with respect to the uncompressed file size) – as shown in Table 8 – then we actually see that ZIP files are, on average, approximately 80-90% the size of LZW compressed files across the four bit-depth levels. It would appear then, that the ZIP algorithm is achieving similar compression improvements over LZW regardless of bit-depth.

Summary with respect to Bit-Depth:

- The average compressed image file size appears to vary with bit-depth, with compression rates decreasing as bit-depth increases (for 1, 8 and 24-bit images).
- 48-bit images appear to achieve better compression than 8 and 24-bit images, with compressed file sizes between 22% and 48% (of the uncompressed size).
- 1-bit images are capable of being heavily compressed down to 0.5% the uncompressed file size using ZIP. The amount of compression achieved appears correlated to the amount of average image entropy in the file.
- ZIP compressed files are approximately ~84% the size of LZW compressed files, across all bit-depth levels.

4.2.3 *Effect of Software Library*

As previously mentioned, both LibTiff and ImageMagick generate average compressed file sizes smaller than the original uncompressed file, regardless of compression algorithm applied. However, as can be seen in Table 7, there is variation between the compression performance of the software utilities for similar bit-depths and compression algorithms. Notably, with the exception of 1-bit LZW compressed images (which is in itself a tiny percentage difference anyway), ImageMagick generates smaller average compressed file sizes than LibTiff. Such an effect is more predominant across the 8 to 48-bit colour depths, irrespective of the compression algorithm used. It is also somewhat true when considered on a file-by-file basis.

Table 9 shows the count of files compressed using ImageMagick which have a file size smaller, larger or the same as those compressed using LibTiff. Specifically these results highlight that ImageMagick generally generates smaller files than LibTiff across all bit-depths – approximately 85% of ImageMagick's LZW files are smaller than LibTiff's; and ~97% of its ZIP files are smaller too.

Table 9: Number of ImageMagick files which are smaller, larger or the same size as those compressed with LibTiff

Alg.	Bit Depth	# Smaller	# Larger	Equal Size
LZW	1	0	56	0
	8	57	0	0
	24	327	18	0
	48	45	0	0
	Total	*429 (85.3%)*	*74 (14.7%)*	*0*
ZIP	1	51	5	0
	8	57	0	0
	24	333	12	0
	48	45	0	0
	Total	*486 (96.6%)*	*17 (3.4%)*	*0*

However there are also occasions when LibTiff fares better and ImageMagick results in larger files; for our sample, this is mainly for 1-bit LZW compressed files, although there are also 5 1-bit (ZIP compressed) images and 30 24-bit (LZW and ZIP) images which are larger.

Analysing the 24-bit "larger" images shows they all come from the same sub-collection of content which have a very large number of unique colours compared to the rest of the sample. Figure 3 plots the difference in compressed file sizes between ImageMagick and LibTiff for 24-bit images. Points above the 0MiB difference line indicate files where the ImageMagick version is larger; points below the 0MiB line indicate files where the LibTiff version is larger. For 24-bit images at least, this plot hints at a (non-linear) correlation between the number of unique colours in an image and the compression performance of ImageMagick (with respect to LibTiff).

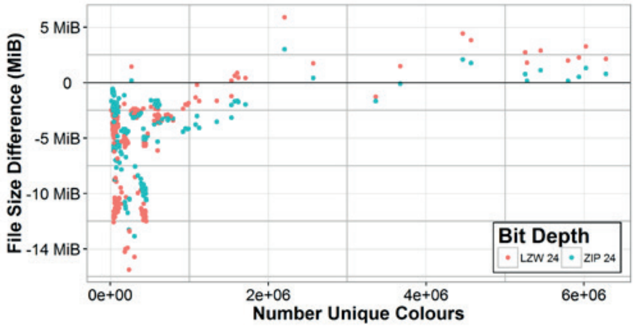


Figure 3: Difference in compressed file size between LibTiff and ImageMagick, with respect to number of unique colours

This all raises an interesting question – if ImageMagick uses the LibTiff library, why should these results differ? Evaluation of the source code shows that for ZIP and LZW compression, ImageMagick uses a slightly higher quality value[6] and sets the TIFF "Predictor" Tag to "horizontal" for RGB and 8/16 bits/sample bilevel/greyscale images[7]. This tag is not set (by default) when using LibTiff directly[8] and requires the user to manually specify the value when compressing an image[9].

The Predictor tag invokes a pre-compression operation which aims to improve compression. The premise for this operation is that subsequent pixels will often be similar in value to their predecessor, especially for continuous-tone images, and so the information content can be reduced by subtracting information already contained in the predecessor pixels. The pixels essentially become the difference from the pixel at the beginning of the row, with many being 0 for continuous-tone images. Compression applied after this can take advantage of lower information content.

ImageMagick's use of this Predictor tag is consistent with our results. With the 24-bit images, the majority of ImageMagick's compressed files are smaller than LibTiff's; where they are larger are for the files which have large numbers of unique colours and high entropy (suggestive of low-continuous-tone). By virtue of how the Predictor differencing works, using this pre-compression operation is unlikely to be as helpful for such files.

These results seemingly contrast with the ZIP to LZW compression ratios shown in Table 8, which show LibTiff offers, on average, slightly better ZIP compression ratio for 24 and 48-bit images than ImageMagick. ZIP compressed 24-bit TIFFs are ~82% (the size of LZW files) for LibTiff versus ~88% for ImageMagick; similarly, 48-bit TIFFs are ~78% versus 80% respectively. Although tempting to think this means LibTiff should offer smaller ZIP files than ImageMagick for these bit-depths, it should be remembered that LibTiff's LZW compression algorithm generates a larger compressed file than ImageMagick's and so the higher compression rates alluded to in Table 8 (with respect to LZW) do not translate to smaller ZIP images. Ultimately ImageMagick generates a smaller average ZIP compressed file than LibTiff, regardless of bit-depth (Table 7).

[6] ImageMagick uses a default quality value of 7, compared to LibTiff's 6; the higher the value, the better the compression.

[7] ImageMagick 6.6.9-5: coders/tiff.c, line 2903 and 2929.

[8] LibTiff 3.9.5: tools/tiffcp.c, line 693

[9] E.g. 'tiffcp –c lzw:2' sets the Predictor tag to 2 (Horizontal)

Finally, whilst the evidence suggests both libraries generate average file sizes less than the original, it also shows that both libraries exhibit cases where LZW compressed files are actually larger than their uncompressed counterparts (see Table 5). Specifically, for our sample LibTiff has over double the occurrences of "larger than original" LZW compressed files than ImageMagick. As previously explained, this is most likely due to limitations with the dictionary based encoding approach used in LZW; however it is also suggestive of implementation differences between the LibTiff and ImageMagick LZW algorithms, such as from the use of the Predictor tag (which favours ImageMagick).

Summary with respect to Software Library

- ImageMagick generates a smaller *average* compressed file size than LibTiff, regardless of compression algorithm.
- Across all bit-depths, our results suggest that:
  - ~85% of ImageMagick's LZW compressed files are smaller than LibTiff's; and,
  - ~97% of ImageMagick's ZIP compressed files are smaller than LibTiff's.
- Some evidence to suggest a correlation between the number of unique colours and whether ImageMagick's compressed files are larger. Further investigation is needed though.
- ImageMagick sets the TIFF "Predictor" tag for RGB and 8/16 bits/sample greyscale images, which could explain its superior compression performance on more continuous-tone images. Further investigation is needed.
- LibTiff appears to offer a slightly better ZIP to LZW compression ratio for 24 and 48-bit images, compared to ImageMagick.
- LibTiff appears more likely to generate LZW compressed files which are larger than the uncompressed file, compared to ImageMagick.

4.3 LZW/ZIP Compressed vs. Group 4 Compressed File Sizes

Whilst the focus of this paper is on the application of ZIP and LZW compression to TIFF files, given we have a subset of files initially Group 4 compressed, it is worth considering how these compare to ZIP and LZW compression. Throughout this discussion it should be kept in mind that Group 4 compression applies to bitonal (1-bit) images only – as such, ImageMagick will not set the Predictor tag.

Table 10 shows the ratio of LZW and ZIP compressed file sizes to the original Group 4 compressed file sizes. As can be seen, compared against TIFFs already compressed using the Group 4 algorithm, LZW and ZIP compressed files are on average, overwhelmingly larger than the originals, with LZW files averaging more than 3 times – and up to 50 times – the Group 4 size, and ZIP files averaging over twice the size – and up to 11 times the size – of the original Group 4 TIFF.

Table 10: Ratio of LZW/ZIP compressed TIFF file sizes as a percentage of the original Group 4 compressed TIFF file sizes (to 1 d.p.)

Library	Alg.	Min	Mean	S.D.	Max
ImageMagick	LZW	69.1%	403.9%	906.6%	5012.2%
	ZIP	55.5%	207.9%	190.8%	1122.5%
LibTiff	LZW	68.2%	366.2%	759.5%	4219.6%
	ZIP	54.6%	212.7%	185.2%	1092.3%

Oddly, these results indicate a difference in resulting file sizes despite the fact that both software libraries do not set the Predictor tag. ImageMagick does use a slightly higher quality setting, which may possibly account for the slightly better ZIP compression (208% vs 213%), however this is not shared in the LZW results (404% vs 366% respectively). It is possible other changes, such as additional tags/metadata, may cause more significant variations in the file sizes seen; further investigation is required.

Table 11: Number of LZW/ZIP compressed files with sizes less than or greater than Group 4 compressed

Library	Alg.	No. LZW/ZIP File size	
		< Group 4	> Group 4
ImageMagick	LZW	11 (19.6%)	45 (80.4%)
	ZIP	12 (21.4%)	44 (78.6%)
LibTiff	LZW	12 (21.4%)	44 (78.6%)
	ZIP	12 (21.4%)	44 (78.6%)

Minimum ratios in Table 10 show that some LZW and ZIP files do compress better than Group 4. Table 11 indicates this is ~20% of such files, consistent across both libraries and algorithms; however a larger sample ideally needs to be tested.

In an effort to understand why certain files compress better and others worse, the LZW/ZIP to Group 4 compressed file size ratio was plotted against the average image entropy – see Figure 4[10]. Recall that entropy is a quality indicating how "busy" an image is, and that higher entropy images should, in theory, be more difficult to compress effectively. Figure 4 suggests that for 1-bit images those with higher entropy are more readily compressible with LZW and ZIP than with Group 4 compression; put another way, Group 4 compression degrades as image entropy increases, which is exactly the result seen earlier in Figure 2.

For our sample there is an outlier file around the 0.25 entropy value (compressing better with ZIP) which goes against our observation, and so further analysis on a larger sample is ideally required before a definitive correlation can be determined.

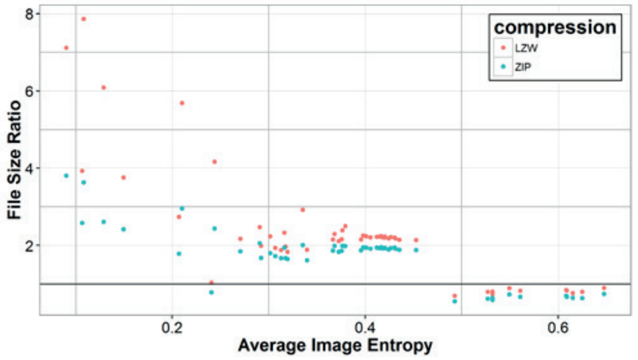


Figure 4: Ratio of LZW/ZIP file size with respect to the Group 4 compressed image file size against the average image entropy (points for 0 entropy not shown)

Summary:

- Group 4 compression applied to 1-bit TIFFs is, on average, at least 2x more effective than ZIP compression, and at least 3x more effective than LZW compression (in terms of generated file size)
- Approximately 20% of 1-bit files compress better with LZW or ZIP algorithms over Group 4.
- There may be a correlation between files with higher entropy values and their ability to be more effectively compressed with LZW or ZIP rather than Group 4 compression.

[10] note: the points for 0 entropy are not shown as they skew the y-axis making the results hard to interpret; these points have a ratio >0

5. WHAT DOES THIS MEAN IN PRACTICE?

This section aims to give practical advice for the application of LZW or ZIP compression based on the previously mentioned findings. The driver for this advice is the reduction of associated file storage, leading to the ability to store more within a defined storage system. Other long-term preservation issues – such as arising from the use of non-baseline TIFF tags or compression algorithms – are not, but should be considered before application of this advice in a production preservation system. Readers should also bear in mind that our observations are obtained from a relatively small sample set, and ideally require further testing.

5.1 For 1-bit TIFFs

Should 1-bit TIFFs be compressed with Group 4, LZW or ZIP to reduce total storage?

Should existing Group 4 compressed TIFFs be decompressed and recompressed with LZW or ZIP to better utilise existing space?

Our analysis indicates that for those TIFFs in the sample which were originally Group 4 compressed (1 bit/pixel, bitonal TIFFs), when decompressed and then recompressed using LZW or ZIP, the original compression was on average more effective than either LZW or ZIP.

This can also be seen in Table 12 below, which compares the mean average and aggregate file sizes for all 1-bit files in our sample in their uncompressed, Group 4 compressed, LZW compressed and ZIP compressed forms. Specifically it indicates that the total space required for the Group 4 compressed sample of 1-bit files is less than for the other compression techniques, regardless of library.

Table 12: Mean average and total size of 1-bit compressed files, by compression algorithm (all to 1 d.p.)

	Alg.	Mean File Size	Total Sample Size
Original	None	940.7 KiB	52677.6 KiB
	Group4	101.4 KiB	5676.5 KiB
ImageMagick	LZW	142.1 KiB	7957.8 KiB
	ZIP	115.1 KiB	6446.1 KiB
LibTiff	LZW	138.9 KiB	7776.2 KiB
	ZIP	118.8 KiB	6653.2 KiB

In essence, the data from our sample suggests that existing 1-bit Group 4 compressed images should be left alone in order to utilize storage space efficiently; uncompressed 1-bit images should be compressed to Group 4.

We did find that approximately 20% of 1-bit files did compress better with LZW or ZIP, hinting at the possibility of selectively encoding 1-bit TIFFs with the most appropriate algorithm to achieve further aggregate space savings. Without conclusive evidence on how to select the "most appropriate" algorithm however, this will probably result in a trial and error approach. It would also diversify the compression profile of TIFFs in a collection.

These results are based on the subset of 1-bit/pixel TIFF images which were originally Group 4 compressed. It is likely that these findings may be extended to other bitonal images, however it should be noted that we have not performed Group 4 compression, and so it is unclear if, and how, these results will be affected through performing such compression with either ImageMagick or LibTiff; further testing would be required in order to formally confirm these conclusion.

5.2 For 8, 24 and 48-bit TIFF images

Should 8, 24 and 48-bit TIFFs be compressed with LZW or ZIP to most effectively reduce total storage?

Which software utility should be used to reduce total storage?

Examination of the file compression ratios has shown that, for our sample at least, ZIP compression is uniformly superior to LZW compression in terms of the degree of file size reduction, irrespective of bit-depth or software library.

Table 13 documents the total space requirements for each bit-depth of our tested sample compressed with each algorithm by each library. This illustrates, particular for 24-bit images, the storage savings achievable through use of ZIP compression over LZW. It also illustrates a preference for using ImageMagick over LibTiff.

While there are some cases where the difference in effectiveness between LZW and ZIP compression is small, there are no examples in this analysis where the ZIP compressed file was larger than the corresponding LZW compressed file. Furthermore, ZIP compression has not caused an overall increase in file size for any images in this sample, which is the case for LZW compression, particularly on 24-bit images.

From a practical perspective, the data from our sample suggests that 8, 24 and 48-bit TIFF images should ZIP compressed using ImageMagick, in order to reduce overall storage space.

Table 13: Mean average and total size of 8-, 24- and 48- bit compressed files, by compression algorithm (all to 1 d.p.)

	Alg.	Bit Depth	Mean File Size	Total Sample Size
Original	None	8	21.6 MiB	1.2 GiB
		24	24.7 MiB	8.3 GiB
		48	36.4 MiB	1.6 GiB
ImageMagick	LZW	8	13.3 MiB	0.7 GiB
		24	16.4 MiB	5.5 GiB
		48	9.5 MiB	0.4 GiB
	ZIP	8	11.4 MiB	0.6 GiB
		24	13.9 MiB	4.7 GiB
		48	7.6 MiB	0.3 GiB
LibTiff	LZW	8	17.0 MiB	0.9 GiB
		24	21.2 MiB	7.2 GiB
		48	16.8 MiB	0.7 GiB
	ZIP	8	15.2 MiB	0.8 GiB
		24	17.5 MiB	5.9 GiB
		48	12.8 MiB	0.6 GiB

Through plotting, we did find that the number of unique colours in 24-bit images appears to suggest a correlation with whether ImageMagick compressed files were larger (than LibTiff's). Although further investigation is needed, this may present a mechanism for selectively encoding 24-bit images using the appropriate library, in order to achieve optimal storage reductions.

5.3 Optimal vs. Recommended Compression

How much total disk space could be saved by using the most efficient compression per file compared to the recommended?

Sections 5.1 and 5.2 presented recommendations, based on evidence from our tested sample, for which compression to apply and what software to use, for each bit-depth of image. Namely:

- 1-bit images: Group 4 compress
- All others: ZIP compress with ImageMagick

It was acknowledged however, that for some files the recommendations were suboptimal (within the bounds of our analysis) with respect to compressed file size. Specifically, some 1-bit images compressed better with ZIP/LZW than Group 4, and some 24-bit images compressed better with LibTiff rather than ImageMagick. If it were feasible to be selective over the compression approach – library and algorithm combination – how much storage would be saved?

Table 14 shows the total storage requirements for the original sample set (i.e. Group 4 compressed 1-bit images; all others uncompressed), plus the storage needs if the recommended or optimal compression approaches were used. It also includes storage figures for an alternative compression approach using Group 4 for 1-bit images and LibTiff ZIP compression (with default settings) for all others.

Table 14: Total sample sizes (in MiB) achieved from original, optimal, recommended and alternative approaches (to 1 d.p.)

	Sample Sizes			
Bit Depth	Original	Optimal	Recommended	Alternative
1	5.5	4.3	5.5	5.5
8	1231.1	651.7	651.7	865.7
24	8512.3	4789.8	4801.5	6039.1
48	1636.4	341.1	341.1	578.0
Total	***11385.3***	***5787.0***	***5799.8***	***7488.3***

In total, there is approximately 13MiB saved from using the optimal approach as opposed to the recommended. Considering the average compressed file sizes presented in Table 13, this equates roughly to an ability to store 1 extra compressed image (out of the ~500 sample).

More generally, the recommended approach has led to an approximate 50% reduction in total file size over the original sample.

As way of example of how the software library employed can have an effect, the alternative compression approach – which use LibTiff ZIP compression instead of ImageMagick's – requires nearly 30% more storage than the recommended approach.

6. CONCLUSIONS

This paper focused on comparing the relative effectiveness of two lossless compression algorithms - LZW and ZIP - on a collection of TIFF files, with the aim of reducing the overall storage needs for the collection. Two software utilities were tested (using default settings) – ImageMagick and LibTiff – to investigate the impact the software choice has on achievable file sizes.

Group 4 compression was found, on average, to be superior to either LZW or ZIP compression when applied to 1-bit bitonal images by at least a factor of 2. Despite this, approximately 20% of our sample of 1-bit images did compress better (on an individual level) with LZW and ZIP. Investigation found some evidence to suggest that the effectiveness of Group 4 correlates (inversely) with the amount of entropy in an image – i.e. "busier" images appear to compress less. However, with only 56 1-bit images in the sample, testing of a larger set would be needed to confirm this.

The ZIP algorithm was found to be superior in effectiveness to LZW for all images in the sample, always generating compressed files smaller than the uncompressed and the LZW TIFFs. In contrast, LZW compression, when applied to the 8-bit and 24-bit images in the sample, occasionally resulted in an increase in file size (from the uncompressed form). This occurred more often when using LibTiff.

The effectiveness of both ZIP and LZW compression algorithms varied with image colour depth, with compression rates decreasing as bit-depth increased (up to 24-bit). 48-bit images seem to buck this trend, achieving better compression rates than 8 and 24-bit images. For specific compression rates see Table 7.

ImageMagick was found to generate smaller average compressed files than LibTiff, with ~85% of its LZW and ~97% of its ZIP compressed files being smaller. Analysis showed that ImageMagick uses the same LibTiff libraries, prompting questions as to why the results should vary so much. Deeper investigation indicated that ImageMagick sets the TIFF 'Predictor' extension tag which enhances LZW/ZIP compression for certain images, offering a probable explanation for the difference, but one that requires further analysis. Theoretically, similar levels of compression should be achievable using LibTiff by setting this tag (no analysis has been performed to confirm this); however based on these results, ImageMagick will perform better by default.

Taking these observations into account, in order to reduce storage space effectively, the following recommendations are suggested:

- For 1-bit images, compress with Group 4
- For all others, ZIP compress with ImageMagick.

For our tested sample, these recommendations result in an approximate storage saving of 50% across the entire collection. It may be possible to reduce the overall storage for a collection further by selecting the most appropriate compression approach on a file-by-file basis; however there is no clear guidance on how to select the best compression approach for any given file, and the overall storage reduction across a collection appears minimal.

6.1 Caveats and Future Work

The figures in this paper should be interpreted with the size of the sample in mind. How these results compare to those obtained from much larger samples remains to be seen, and would be useful further work. In particular, it would be useful to test on a sample set that encompasses larger sub-collections, i.e. a sample with larger numbers of 1-bit, 8-bit and 48-bit images.

Given the connection and results variation between software utilities shown, evaluating the performance of these libraries using the same settings would be of benefit, for example, comparing LibTiff with the Predictor tag set to 'Horizontal' and a quality level of 7 (as per ImageMagick).

It should also be borne in mind that no Group 4 compression was undertaken. 1-bit files were already Group 4 compressed in the sample, and these were used as is. An obvious enhancement to these experiments would be to start with uncompressed TIFFs and Group 4 compress them using LibTiff and ImageMagick.

Compression of a TIFF file is applied to the image payload only. Whilst it might be expected that this would be the only source of change in a file when compressing, additional changes also occur in the tagged metadata portions of the file to describe the compression. Furthermore, additional metadata, particularly that associated with the Adobe Photoshop's "Image Source Data" tag (# 37724), which captures layering information, appears to be removed during compression. Such changes to the tagged metadata are included in the file sizes presented in this paper. Therefore the change in file size represents the complete change to the file, and not just the change to the image pixel data. Consideration should be given as to whether this presents vital information that must be preserved, and therefore whether compression is appropriate.

Similarly, this paper does not address other long-term preservation issues with the use of TIFFs, non-baseline tags and compression. Robustness of compressed TIFF formats towards bit-errors is not examined, although perhaps mitigated through bit-level preservation. LZW and ZIP compression are both TIFF extensions which do not have to be supported by TIFF readers, as is the Predictor tag. Subsequently, there is a small possibility that compressing TIFFs may make them difficult to render with baseline-compliant-only TIFF readers. Whilst there is currently software (e.g. LibTiff) able to decompress such files, consideration needs to be given to the appropriate preservation practices and documentation required for the software and algorithms involved.

7. REFERENCES

[1] Adobe Photoshop® TIFF Technical Notes. 22 March 2002. http://partners.adobe.com/public/developer/en/tiff/TIFFphotoshop.pdf [Online; cited: 24 Apr 2016]

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

[3] LZW Patent Information. Unisys. 02 June 2009. https://web.archive.org/web/20090602212118/http://www.unisys.com/about__unisys/lzw [Online; cited: 24 Apr 2016]

[4] Mateika, D. Martavicius, R. 2006. *Analysis of the Compression Ratio and Quality in Medical Images*. ISSN 1392-124X. Information Technology and Control, vol. 35, No. 4

[5] Puglia, S., Reed, J., Rhodes, E. 2004. T*echnical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images*. U.S. National Archives and Records Administration (NARA). http://www.archives.gov/preservation/technical/guidelines.pdf [Online; cited: 12 Apr 2016].

[6] Succeed Project, 2014, *D4.1 Recommendations for metadata and data formats for online availability and long-term preservation*. http://www.succeed-project.eu/sites/default/files/deliverables/Succeed_600555_WP4_D4.1_RecommendationsOnFormatsAndStandards_v1.1.pdf [Online; cited: 12 Apr 2016].

[7] *Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files*. 2010. Federal Agencies Digitization Guidelines Initiative (FADGI) – Still Image Working Group. http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf [Online; cited: 12 Apr 2016].

[8] The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants. 2015. Library of Congress. https://www.loc.gov/ndnp/guidelines/NDNP_201618TechNotes.pdf [Online; cited: 12 Apr 2016]

[9] TIFF Revision 6.0. [Online] 3 June 1992. https://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf. [Online; cited: 24 Apr 2016]

[10] Wheatley, P., May, P., Pennock, M., *et al.* 2015. *TIFF Format Preservation Assessment*. http://wiki.dpconline.org/images/6/64/TIFF_Assessment_v1.3.pdf [Online; cited: 24 Apr 2016]

Towards a Risk Model for Emulation-based Preservation Strategies: A Case Study from the Software-based Art Domain

Klaus Rechert
University of Freiburg
Hermann-Herder Str. 10
79104 Freiburg, Germany
klaus.rechert@rz.uni-freiburg.de

Patrícia Falcão
Time Based Media
Conservation, Tate
7-14 Mandela Way
London SE1 5SR, U.K.
patricia.falcao@tate.org.uk

Tom Ensom
Department of Digital
Humanities, King's College
Drury Lane
London, WC2B 5RL, U.K.
thomas.ensom@kcl.ac.uk

ABSTRACT

Virtualization and emulation provide the technical basis for a potential preservation strategy to keep performing digital objects accessible, despite obsolescence of their original technical platform. By simulating the presence of physical components, in the form of virtual hardware, it is possible to maintain the software environments needed to run original software.

In this paper we describe a conceptual model to analyse and document the hardware and software elements of software-based artworks, in order to then identify dependencies and assess risks. This information is used to select the most appropriate options when creating disk images, and to make decisions about whether an emulation or virtualization platform is more appropriate in supporting these disk images. We conclude with recommendations for ensuring the sustainability of disk images in the medium-to-long-term, and strategies to mitigate the risks related to external dependencies.

Keywords

Software-based Art; Emulation; Preservation Strategy

1. INTRODUCTION

Successful preservation of software-based artworks requires a deep understanding of both the technical aspects that underlie the performance of the software and the properties of the artwork that must be maintained to ensure its authenticity. Only by looking at those two aspects together is it possible to determine the relevant technical strategies to apply.

Software-based artworks are typically understood as artworks for which software is the primary medium. Software usually makes use of source code, as in Richard Rinehart's definition, *"... There is one aspect of digital media that separates them from traditional art media and even other electronic art media; source code."* [15] Sometimes however, software might be created at a level abstracted from the source code itself – through the use of production tools (for example, a WYSIWYG or visual editor).

Currently, Tate's collection includes only ten software-based artworks, a small number even in the Museum context. However, these artworks have been produced by markedly different artists and programmers, over a period of ten years; resulting in a wide variety both in the functions performed by the software and in the hardware and software systems used. The software is often custom-built for a specific artwork, so there are no standards as to how it is built or documented. The unique nature of the technologies used means that individual preservation plans are essential, and for some of the artworks emulation has already proven to deliver satisfactory results. However, the expected growth of the collection and the related practical and economic constraints highlight the importance of identifying common features and developing strategies that can be applied consistently, so as to make the preservation of the artworks and related artefacts sustainable.

The proposal in this paper aims at keeping the digital artefacts required to perform an artwork available and accessible by preserving the technical platform they were developed for. These platforms, partly physical (hardware) and partly non-physical (software), are superseded by new platforms every five to ten years. Once the hardware parts are out of production, the software parts also become inaccessible and old platforms disappear from the market and general use. To keep born-digital objects accessible, a promising approach is to focus on keeping the hardware platform alive by simulating the presence of the physical parts through virtual hardware. Virtualization and emulation are able to provide the technical basis for a potential preservation strategy to keep performing digital objects accessible, despite obsolescence of their original technical platform. Virtualization and emulation are proposed as generic technical preservation strategies, which can be shared among similar artefacts. The process of creating an emulated version of an artwork has the advantage of highlighting preservation risks posed by a constantly changing technical environment. It is also the moment to evaluate specific external dependencies the digital object may have and identify possible solutions or steps that can be taken to reduce the impact of the loss of these dependencies.

We describe a methodology for evaluating whether emulation or virtualization can or should be applied to the digital artefacts that make up a software-based artwork. The methodology combines an assessment of risks for preservation, a proposal for best-practice when migrating between physical and virtual environments, and consideration of how to maintain virtual machines in the long-term. The final section addresses the fact that emulators themselves become obsolete. It discusses ways in which dependency on a particular emulation platform can be reduced and how best to migrate to a virtual hardware environment in order to facilitate long-term access to the software. The processes described in this paper have particular potential for implementation in art museums, but may also be broadly relevant to other types of software collection.

2. RELATED WORK

Emulation as a preservation strategy has been discussed for 20 years, an early example being Jeff Rothenberg's paper from 1995, "Ensuring the Longevity of Digital Documents" [17]. These early ideas were later reframed in an art context by Richard Rinehart's work for Rhizome [14]. The term emulation was also used in various other projects in art conservation, but often in a wider sense, to refer to the re-coding of a work or changing the technologies used to instantiate a work [19, 6].

Hardware virtualisation was first specifically proposed as an approach in the context of software-based art by Tabea Lurk in 2008 [9], and positioned as a potential conservation tool. The SCART project investigated the use of emulation, virtualization and re-coding for Mondophrenetic™.[1] Today, however, emulation is still neither common practice nor has it evolved from singular project-based experiments.

In the last years, further research by various institutions has resulted in progress regarding usability and scalability of emulation and virtualization for the preservation of complex digital objects. Some of the most significant of these projects were the Olive Archive [10], the bwFLA Emulation as a Service [13] and the Internet Archive's Emularity[2]. While all three approaches greatly reduced technical hurdles and seem to be ready for broader adaptation and usage within the preservation community [16], there is still a lack of generic emulation-based preservation strategies. Recent research on software-based art preservation has mostly focused on CD-ROMs [5, 2, 3].

The notion of significant properties has been examined for complex digital objects of various kinds, including software, and these studies have included discussion of an artefact's technical features and dependencies [7]. However identifying and classifying these kinds of properties can be challenging due to, *"The diffuse nature of software-based artworks and the systems of which they are made, means that obsolescence is often more difficult to monitor than in traditional time-based media works of art and the risk of failure in these works is harder to predict."* [8] Further to that, for artworks, the concept of significant property must extend beyond the properties of the digital objects themselves. It must include other elements that influence the experience and understanding of an artwork, such as the spatial parameters of display in the gallery space [8].

With software-based artworks now making their way into the collections of art museums typically associated with more traditional media, there is a pressing need to address the challenges of preserving these works and to develop the associated skills among practitioners. Best practices however, are not yet available, or indeed, arrived at with any real consensus. It is hoped that this paper will be a useful contribution to this developing area and help provoke a movement toward agreeing best practices among the community.

3. TECHNICAL CHARACTERIZATION

At first glance a digital artefact consists of a set of byte streams, e.g. binary files. Keeping these accessible in the long-term (i.e. being able to retrieve exactly the same byte stream as originally stored) poses some risks, but from today's perspective is a manageable task supported by established procedures and tools. In contrast, keeping a digital artefact's experienceable features accessible is a much harder task, since the artefact needs to be rendered, or performed (and possibly requires interaction with viewers/visitors) and depends on a suitable technical environment to do so. To ensure the long-term availability of a computer platform's hardware (e.g. to render a (preserved) digital artefact) emulation and virtualization can be considered as potential access and preservation strategies. In order to prepare an emulation-based preservation plan, a detailed technical analysis of the digital artefact and its technical environment is required, to uncover explicit but also implicit dependencies as well as determine present and future risk-factors.

3.1 Software Layer

Many digital artefacts are not self-contained. They do not only require hardware, but also additional software to be rendered. Therefore, one part of an artefact's technical environment represents a software runtime – *the software (rendering) environment*. A minimal software environment is typically an installed and configured operating system, but in most real-world scenarios a more complex software environment with additional software installed and configured is required. When acquiring an artefact, its software environment needs to be assessed.

3.1.1 *Interfaces between Hardware and Software*

Operating systems (OS) play an important role in software environments, as they typically provide a hardware abstraction layer, for instance, any application is able to connect to the internet without knowing technical details about the hardware used. This abstraction is usually achieved through *technical interfaces* – the so called hardware abstraction layer.

The technical interfaces between an operating system and hardware have two sides: the top-side (OS-side) unifies usage of different hardware components (e.g. sound card, network card, graphic cards etc.) and the bottom part (hardware-side) operates the hardware in (vendor-) specific ways. The connection between top- and bottom interfaces are implemented as hardware drivers (sometimes as BIOS or ROMs extension[3]).

Through the use of OS hardware abstraction, software dependencies on physical hardware components are usually unnecessary. Software artefacts then pose only abstract hardware dependencies (e.g. the minimal screen resolution, sound and network support etc.). This approach has greatly simplified software development and improved the compatibility of software with a wide spectrum of different computer setups, since most artefacts and software dependencies typically use operating systems to interact with hardware.

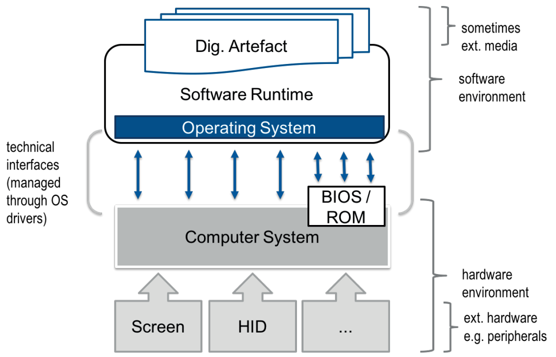


Figure 1: Rendering environment of a dig. artefact.

Some digital artefacts, however, have no software dependencies and are able to interact with hardware components directly. However, these cases are rather rare (at least for typical computer setups) and usually associated with specific hardware – for example, game consoles, arcade machines, robots or similar special purpose machinery. There are also some rare cases where an operating system is used but the artefact also relies on direct access to a specific hardware resource.

3.2 Hardware Layer

The hardware layer connects the digital artefact (and its software runtime) with the physical world. Hardware components, as considered in this paper, can be classified into two classes: a 'machine' with built-in hardware (e.g. a computer, phone, tablet etc.), and external hardware components connected to this machine. For the purpose of this paper the hardware's properties are only relevant in so far as they influences the options for emulation or virtualization. The focus of this section then, is on the hardware characteristics to document when considering its use for emulation or virtualization purposes.

When an artwork is acquired it is important to analyze and describe the hardware used by a digital artefact, as this will help to define the technical environment required for that digital artefact to be rendered. The level of detail needed to describe the hardware depends on the characteristics of the software environment where the digital artefact is run.

3.2.1 *Virtual Hardware*

Any computational (binary) operation can be implemented in hardware (i.e. hard-wiring operations for a specific purpose as a machine or machine component) or in software (i.e. translating a set of complex logic operations by compiling source code into instructions for a generic machine – e.g. any generic CPU). Both implementation options are in theory equivalent, however operations implemented in hardware are usually by magnitudes faster compared to a pure software implementation. This equivalence allows, however, the replication of any outdated hardware component in software and the exposing of its functionality using contemporary hardware. Hence, the *Computer System* block, depicted in Fig. 1 can be implemented either as physical or virtual hardware.

There are currently two generic technical options to replace outdated hardware: virtualization and emulation. These two technologies are not mutually exclusive, and in fact share many concepts.

Virtualization. Virtualisation is a concept and a technical tool to abstract (*virtualize*) hardware resources, such that a so called virtual machine (VM) (also called guest) is not interacting with the physical hardware directly. Instead a hypervisor (also called host) provides a virtual hardware interface for guests, usually providing access to a unified set of (mostly emulated) hardware, regardless of the machine's hardware configuration the host is running on. A virtual machine is still able to utilize performance advantages of real hardware, in particular (but not restricted to) using the host's CPU. The hypervisor is in charge of enforcing rules as to how a virtual machine is able to use the host's CPU (i.e. restricting access to specific, sensitive instructions – for instance, preventing a VM accessing the host's memory), such that a VM is unable to takeover the physical machine or interfere with other VMs running on the same host. Modern computer architectures have built-in hardware features (e.g. dedicated CPU and memory-management features) to support virtualization, implementing parts of the hypervisor in hardware and thus reducing the virtualization overhead as well as the complexity of the host system (software hypervisor).

Hardware components available to guest VMs are either fully emulated or, to improve performance by eliminating most of the emulation overhead, paravirtualized [18]. Paravirtualized hardware offers (almost) direct and efficient access to the host's hardware, typically (but not restricted to) network cards and storage controllers (e.g. disk access). When using paravirtualized hardware, a driver specific to the virtualization system needs to be installed within the guest system. In contrast, when using fully emulated hardware components, the guest system is able to use the same driver code as if it were using a physical hardware component.

Emulation. An emulator (usually) implements a specific outdated computer system, primarily a CPU architecture, interpreting the outdated machine's instructions and translating these to equivalent instructions of the current host system. This however, is not sufficient to run or execute even the simpler applications. There is additional hardware emulation required to attach storage devices (e.g. a bus to a floppy or hard-drive controller), a memory management unit (MMU), video and audio output, network interfaces, and interfaces for interaction with users. In contrast to virtualization, emulation implements a complete computer system solely in software. Therefore, an emulated system is independent of the current computer architecture, e.g. we are able to run a Motorola 68k emulator (e.g. to boot MacOS System 7) on a current Intel-based Windows PC.

3.2.2 *Virtualization or Emulation?*

The main difference between emulation and virtualization is the reliance on contemporary hardware (or the lack thereof). Virtualization relies on and utilizes real hardware for performance reasons as it offers typically (almost) native execution speed, but has restricted platform support. By

[1] Mondophrenetic™ a work made by Herman Asselberghs, Els Opsomer and Rony Vissers in 2001, https://www.scart.be/?q=en/content/case-study-report-mondophrenetic-2000-herman-asselberghs-els-opsomer-rony-vissers-0

[2] http://digitize.archiveteam.org/index.php/Internet_Archive_Emulation (online, 4/22/16)

[3] Apple Macintosh Computers are a good example for relying on ROMs

definition, virtualizers (such as VirtualBox and VMWare), are only able to run Intel-based x86 VMs, whereas emulators cover almost any technical platform, in particular obsolete ones. Furthermore, the close ties to today's computer platforms restrict a virtualized machine's longevity, particularly if the virtual machine relies on contemporary (paravirtualized) hardware components. To support paravirtualized hardware, the VM (and the virtualization technology) not only rely on VM-specific drivers installed in guest systems, but these drivers also expect appropriate support from the host OS, typically as a host-OS kernel extension to support interaction with the host's hardware directly. Any major OS-kernel upgrade (e.g. changes to internal kernel interfaces) requires an upgrade of the virtualizer too, and therefore the longevity of the virtual machine guest depends also on vendor or community supporting current operating systems.

As all hardware components (and respectively their low-level interfaces to be used by drivers) of a computer system have to be re-implemented in software, and the availability of drivers for old operating systems is crucial, only a small set of emulated hardware components are provided as virtual or emulated hardware. Typically, emulator developers focused on hardware in widespread use and with good driver support, e.g. Soundblaster 16/32 for soundcards and Intel's E10/100/1000 for network cards. In practice there is a significant overlap between virtualizer and emulators, with both supporting a similar set of emulated hardware components, a useful property for a mixed preservation strategy. Hence, for digital preservation and related tasks, one should avoid extensions or drivers specific of a certain virtualizer or emulator. If possible, drivers originally published by hardware vendors should be used, since using the original driver also verifies (at least partly) correctness and completeness in the emulated hardware. Furthermore, by using emulated standard hardware and their drivers, both the effort and risk of migrating a virtualized system to an emulated one is reduced. Contemporary emulators require a specific contemporary host system, which is to say that emulators are normal software components with specific requirements regarding their software and (abstract) hardware environment. However, the guest systems running on emulated hardware are usually not specifically configured for a certain emulator (e.g. using original hardware drivers). Hence, migrating an emulated guest to a new emulator of the same hardware architecture, will require little or ideally no adaptation of the system.

To summarize, using virtualization technology can be a useful addition to a general emulation strategy, in particular if performance matters. Even though a virtualization solution can not be considered a long-term solution, if carefully configured (e.g. avoiding paravirtualized drivers) the effort required to migrate a system to a new virtualizer or emulator is lowered. Furthermore, having two similar systems at hand (e.g. VirtualBox for virtualization and QEMU for emulation) offers the option to pursue a two-track strategy, and in particular allows to practice system migrations between two virtual hardware stacks.

3.3 Conceptual Layers

Based on the aforementioned structural view and discussion a (minimal) technical description of a digital artefact can be divided into three conceptual layers:

1. *Artefact Description & Configuration* This layer conceptually captures the technical description of the object, in particular the technical properties of the artefact itself and its specific (artefact-specific) configurations and settings.

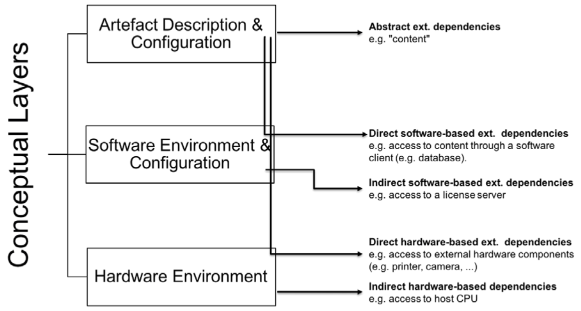

Conceptual Layers
Artefact Description & Configuration
Software Environment & Configuration
Hardware Environment
Abstract ext. dependencies
e.g. "content"
Direct software-based ext. dependencies
e.g. access to content through a software client (e.g. database).
Indirect software-based ext. dependencies
e.g. access to a license server
Direct hardware-based ext. dependencies
e.g. access to external hardware components (e.g. printer, camera, ...)
Indirect hardware-based dependencies
e.g. access to host CPU

Figure 2: Characterization of external dependencies derived from conceptual layers.

2. *Software Environment & Configuration* This layer conceptually captures all installed software components and applications, including the operating system (if present). Furthermore, it may capture the configuration and settings of individual software components and the operating system.

3. *Hardware Environment* This layer conceptually captures the hardware components used by both upper layers.

3.4 Characterization of External Dependencies

A digital artefact requires more than an isolated technical environment (consisting of data and/or extracted disk images) and a computer system or installation to be rendered. For this reason, external dependencies need to be determined, characterized and documented.

The three logical layers together describe the technical environment and characteristics of a digital artefact. In each layer, individual components may depend on functionality or data not available within the local setup (*direct* external dependencies). Additionally, there may be *indirect* external dependencies. For instance, a direct software dependency originating from the artefact layer may itself have external dependencies.

From the three conceptual layers we can derive the following five types of external dependency:

1. *Abstract external dependency:* Abstract dependencies are posed directly by the digital artefact. These dependencies are abstract in that they do not rely explicitly on a specific software or hardware component. For instance, an artefact's performance might depend on access to some kind of data source stored at an external site, but does not rely on specific software to retrieve or modify the data (e.g. the data is directly accessible through the local file system).

2. *Direct software-based external dependency:* To access external data and/or functionality the artefact requires additional software. For instance, a specific client software is required to connect to a remote database and retrieve data.

3. *Indirect software-based external dependency:* This type of external dependency is not posed by the artefact directly but by another of the artefact's software dependencies. It is therefore called an indirect software dependency. For instance, database client software might require access to a license server to function.

4. *Direct hardware-based external dependency:* The digital artefact requires access to external hardware, such as direct access to a specific printer.

5. *Indirect hardware-based external dependency:* The software environment of the digital artefact requires access to specific hardware, e.g. a software component requires access to a hardware license dongle to function.

3.4.1 Peripherals

An important subset of external dependencies are external (connected) hardware components, which can be seen as direct or indirect hardware-based dependencies. A general characterization of external hardware is beyond the scope of this paper. Instead, this section will focus on the characterization of the communication between a virtualized or emulated machine and external hardware, as well as data protocols used, (i.e. how information is exchanged between software and external hardware). This is the essential information needed when considering the use of emulators to run an artefact.

To (re-)connect external hardware components a physical machine is required. In the case of an emulated or virtualized computer system, the host system needs to be able to connect and interact with external hardware components such as human interface devices (HID) (e.g. mouse and keyboard), printers or other peripherals, e.g. by using a suitable connector or a passive (if electrically compatible) or active (e.g. analogue-to-digital converter) adapter, to provide a compatible connection.

Second, the host operating system needs to provide a software interface for applications to communicate with external hardware. For example, a COM port, the Windows software-side representation of a serial connection used for generic peripheral devices such as printers. If external hardware are accessible through the host OS's interfaces, an emulator (acting as a normal software application) is then able to use this external hardware.

Finally, the emulator needs to provide a virtual hardware interface connected to the host's software interface, visible to and usable by the guest OS. Through all these layers integrity of the data protocols, used to communicate between software within the emulated environment and external hardware, needs to be maintained.

Similarly to the previously discussed built-in hardware components, judging the relevant and controllable risks posed by an external component on the complete installation should be focused on its technical interfaces. Fortunately, the number and type of external technical interfaces is low. Their types are standardized and mostly use general purpose technologies (such as USB, serial, parallel etc.). Some older external components and interfaces aren't supported by emulators anymore, mostly for practical reasons, such as host systems providing no appropriate connectors (e.g. a mobile phone or tablet being used as an emulator host has limited connector options). In these cases, usually an emulated or simulated option is provided (e.g. a GUI console gamepad emulation on a touchscreen device).

4. AN EMULATION-BASED PRESERVATION STRATEGY

In the previous section three conceptual layers describing the technical characteristics of a digital artefact were identified. The structural separation of hardware, software environment and the digital artefact facilitates the evaluation of preservation risk factors and strategies without considering the specificities of all layers.

By choosing a virtualization or emulation strategy the focus of preservation moves from physical objects (*Hardware Environment* layer) to disk images (*Software Environment* layer). Hardware will inevitably suffer from technical and physical decay, and for cost reasons can only be preserved in individual cases (even then eventually failing). The same applies to emulators. In contrast, preserved software environments don't change in their hardware requirements over time and can be considered as constant and stable in their requirements. The goal of this strategy can be described as having software environments (disk images) to *run anywhere* and *run forever*. *Run anywhere* translates into making a complex software installation portable. Most relevant is to create the most generic software environment possible, with regards to hardware dependencies, such that a digital artefact is able to perform outside of its original environment. *Run Forever* can only be achieved if the disk images are maintained over time. For that both the technical interfaces and external dependencies of disk images must be regularly monitored. This can be done either by referring to technical documentation, if available, or by performing periodical tests to assess the functionality of the dependencies. If an interface's functionality breaks or an external dependency becomes unavailable, the disk image (respectively its software environment) or the artefact itself must be adapted so they can perform again in the changed technical environment.

For that to be possible, steps must be taken when creating the disk images to facilitate their longevity. Alongside this there must be careful planning for their obsolescence, or the obsolescence of their technical environment, in particular the emulation or virtualization platform. The first step for a successful emulation strategy is knowing what information and resources are available to support this process. Having the artefact's source code available (enabling the recreation of the artefact on a current machine) allows for a higher technical abstraction level and may also open the door to alternative strategies (e.g. a traditional migration approach). If source code is not available, creating an emulated version of the work may be the only suitable choice. How this emulated version is then created will depend on whether the software can be re-installed and if all the required software dependencies are available and operational. Hence, for an emulation-based preservation strategy the artefact's software environment needs to be determined, in particular its software environment's composition and the technical interfaces to the hardware layer (cf. Section 4.1).

Having a (detailed) software environment description allows to focus preservation planning and preservation action activities on monitoring and managing the technical links between software and hardware environments. While these links are stable in the short term, emulators are also subject to the software life-cycle, and will become technically obsolete at some point. Then, if new emulation software is required, a new (emulated) hardware environment needs to be determined which meets the technical requirements of the software runtime. If the technical interfaces between hardware and software environment are well documented – as a vital part of a software environment description, all affected software environments can be determined, and the search for new emulators can be guided effectively. If no perfect match is found, the required adaptations of the affected software environments can be predicted in an automated way using this same documentation (cf. Section 4.2).

For the remainder of this section we assume that the artefact itself is either already available as a file or bitstream or the artefact is part of the disk image. Furthermore, we assume that the artefact's significant properties have been assessed and the artwork is available for verification purposes.

4.1 Acquiring Software Environments

The task of determining an artefact's software environment, starts with the analysis of the artefact with the goal to produce an accessible, performing setup without relying on the availability of physical hardware components and to gather enough information to support future preservation tasks, i.e. ensuring the artefact's longevity. Depending on the artefact's composition, e.g. individual technical components present – a binary object, its configuration, possibly the object's source code, any kind of documentation and a reference installation in form of a computer system – different options are available to pursue both goals. Furthermore, the levels of documentation depend on a series of factors, primarily if the artist and/or the artist's programmer has supplied any technical details about the software, and whether they are available to answer any questions about the work's technical makeup, runtime process and system configuration. In a first step, all components of an artefact are assessed regarding information about the artefact's technical dependencies, i.e. software, hardware and external dependencies and to support the selection of virtual hardware.

4.1.1 Selecting Virtual Hardware

Emulators and virtualizer provide only a limited selection of supported hardware components. Their capabilities are best described by a list of supported computer systems (e.g. x86 ISA PC or Apple Macintosh Performa) and a list of supported operating systems. Therefore, the most important information to be documented in the original computer system to be preserved is therefore the general hardware architecture (e.g. CPU type, bus architecture or ROM type/version), in order to help choosing an appropriate emulator or virtualizer. The choice of an emulator or virtualizer also requires information about the software environment (e.g. the emulator must support Windows 3.11), as even if an emulator supports a particular computer system, it may not support – or support only partially – an apparently compatible operating system. Detailed hardware information will only rule out incompatible emulated computer platforms. The final decision on a suitable emulator/virtualizer requires that support for both the computer system and the associated software environment (primarily operating system support) are assessed.

A further, more detailed comparison between the identified hardware components and the features of a specific emulator is useful to estimate up-front the work required to migrate the machine's software environment (disk image) to a new hardware environment. A detailed list of hardware components installed provides insights on how the operating system was configured. For example, by comparing a detailed list of hardware components with a list of the hardware supported by an emulator it is possible to predict if a software environment (in form of the computer's disk image) will boot directly using emulated software. If the system is able to boot to a certain point (e.g. a simple 'safe' environment with all non-essential hardware features disabled), built-in operating system tools can be used to adapt the system to the new, emulated hardware environment. These adaptations could involve identifying available hardware or suggesting drivers.

The importance of specific hardware components can only be assessed using information about the artefact and/or its software environment. For instance, if the hardware setup involves a high-end 3D graphics card, its importance to the whole installation can be judged on how the graphics card is used: is the performance of the card a key factor or does the software depend directly (i.e. direct hardware access) or indirectly (i.e. through an abstraction layer such as DirectX or OpenGL) on specific hardware features.

The necessity of incorporating software environment information (to assess the role of hardware components) as part of a digital artefact's technical environment highlights the importance of the technical interfaces between software and hardware environment. These pose high risks to the artefact's functionality.

4.1.2 Workflows

To provide a runtime environment for a digital artefact, any emulation-based preservation strategy is likely to result at some point in managing a set of software environments or virtual disk images containing instances thereof. Disk images may contain the digital artefact (or parts of it), or might be prepared separately from the digital artefact, which is held on separate virtual media (or similar) and to be used with this disk image. Three different workflows/strategies can be applied to the task of image acquisition, depending on practical options and requirements, as well as on information available about the artefact and environment. Fig. 3 illustrates the software environment acquisition process.

Generalization: If the original computer system is available, images of physical media (e.g. a hard disk) can be made. To make these useable in an emulation environment and to facilitate the long-term management of disk images, as a first step it is necessary to *generalise* the technical interfaces between hardware and the lowest software layer (typically the OS, respectively OS hardware drivers and configuration). In case of disk images originating from physical media, generalisation is part of *migrating* a software environment from physical hardware to virtual/emulated hardware. Generalising means that any specific drivers (for instance, from a disk image made from a physical computer) are replaced with drivers supported by the virtualization or emulation platforms in use. Hardware dependencies should be systematically determined and all necessary adaptations kept to a minimum when migrating to virtual / emulated hardware, e.g. to maintain the environment's authenticity.

As part of this process, the choice of emulated hardware and drivers should be consistent, so that for each combination of OS and hardware platform always the same (emulated) hardware component is used. This means that for each emulated hardware platform and all associated software environments there is only a limited number of technical interfaces to be maintained and monitored, and this consequently means that the same migration strategy can be applied to different software environments. For example, ten different physical computers may use ten different video cards, which may each use different drivers with the same functions. By generalising the disk images created from these computers the number of drivers needed for that video card can be reduced, so that instead of ten different drivers only one is needed, and later on only one may need to be replaced (if necessary at all) for migration to another emulator.

As a result the generalisation workflow produces a disk image to be used with a contemporary emulator. Additionally, external and hardware dependencies are uncovered by running the artefact and its software setup in a different virtual environment. In particular, the image's technical interfaces can be documented by running on well understood virtual hardware. To ensure that the significant properties are not affected the initial process of image generalisation should be performed during or after image acquisition, and preferably a comparison between the original and generalised systems should be made.

Rebuilding A second workflow is necessary if either no computer system was available to be imaged or – in order to reduce future preservation risks – a secondary set of software environments (disk images) are desired.

Ideally the configuration of a software environment is known, i.e. available as (machine readable) metadata, such that the software environment can be rebuilt if necessary. If the software environment is not known, an artefact's dependencies may be determined by using its original environment as a reference (e.g. the artist's original computer). This reference environment can be analyzed and the artwork can be isolated from its original environment to be rendered in another technically compatible environment. Either using documentation derived from a reference setup or systematically determined software dependencies (e.g. using tools for analyzing an artefact's technical format or its runtime behavior [4]), a suitable software rendering environment can be remodeled by re-installing and re-configuring an artefact's software environment in an emulated environment.

When re-building environments, a consistent configuration of an operating system is built. For efficiency reasons, a specific operating system on a specific hardware platform is only installed once, any more sophisticated software environments are derived from these base images. Also in this case the initial choice of the system's configuration matters, as it will affect the preservation options of all derived environments. The choices on a software environment's hardware components could be based on emulator support (do other emulators/virtualizers, in particular open source, support this hardware component?), the popularity of the device while in production and available driver support. If a popular and tested open source implementation of this hardware component is available, it seems more likely that future emulators will resort to that implementation instead of implementing a historic hardware component from its specifications. The same applies for the availability of drivers: if a hardware component has been successfully used in emulators even after the component was out of production, it is highly likely that archived versions of drivers of these hardware components remain available.

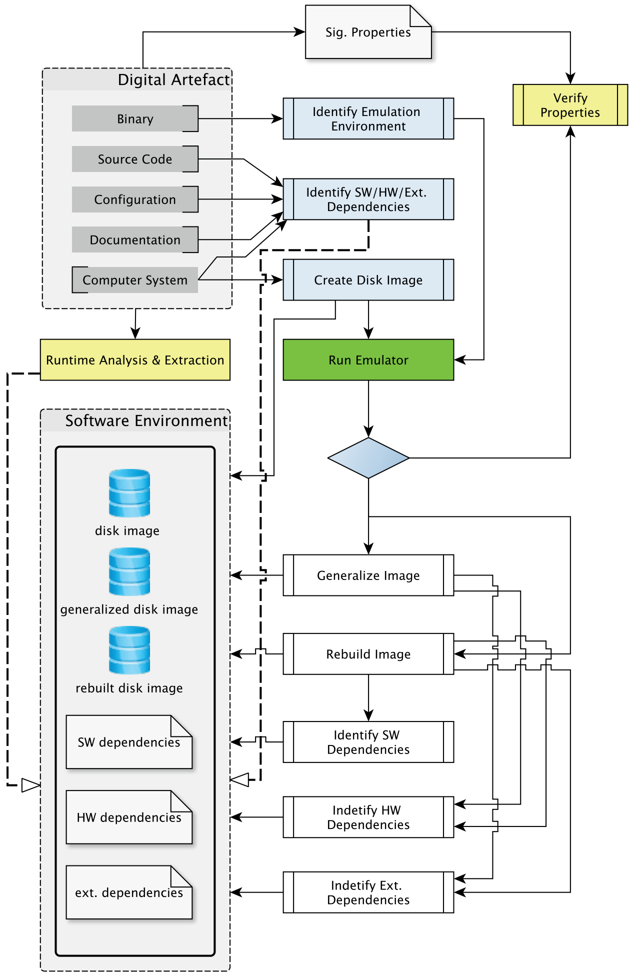


Figure 3: Acquiring software environments

The process of manually rebuilding an artefact's rendering environment can also be used to create verified and machine readable technical metadata. Machine readable installation information (e.g. what type of software is installed), and more importantly the environment's configuration, may be created in an automated way during a structured software environment rebuilding workflow [12].

Pre-built Environments In some cases a complete computer system is not available, i.e. only the (binary) artefact is available without a reference runtime or setup. In this case an alternative is the use of pre-built software environments. These environments resemble typical computer systems of a certain period, e.g. a typical MS Windows 98 installation equipped with popular software, utilities and libraries. In this case a suitable pre-built environment needs to be identified for a given digital artefact and, if necessary, adapted to the artefact's specific requirements. Similar to re-built environments, this approach results in a well documented

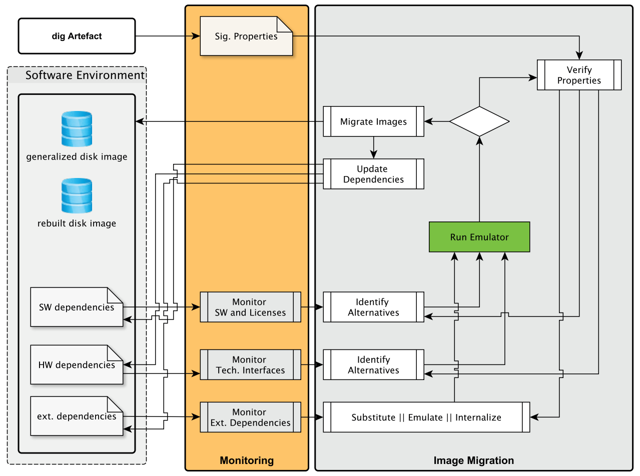


Figure 4: Maintaining software environments

and well understood (base) environment, shared among a set of similar digital artefacts.

4.2 Maintenance and long-term Preservation

The outcome of a software environment acquisition workflow is a disk image (representing an instance of a software environment). Even though the technical file format of the resulting disk image is identical in all strategies (usually a virtual hard-disk, bootable by an emulator) different workflows must be followed for maintenance and long-term preservation. The actual available preservation actions will depend on information about the content and configuration of the disk images, in particular its technical dependencies, rather than on their file format. In order to maintain a software environment's *run forever* property, its technical dependencies require monitoring, particularly:

- monitoring of software components used, including availability of licenses and external dependencies;
- monitoring of existing and emerging emulation and virtualization technologies for continued support of identified technical interfaces;
- monitoring of external dependencies.

Through monitoring technical dependencies, the identification of an imminent risk of obsolescence may indicate the need to migrate a disk image. This strategy then becomes very similar to the one applied to simpler digital objects such as spreadsheets or word processor files. In general, to implement a migration strategy, objects are monitored regarding their technical support (and therefore proximity to technical obsolescence) and migrated to a new, more sustainable or current format if required. Ideally all the significant properties of these objects' are preserved in the process. These interfaces may break if an emulator drops support for specific hardware (e.g. emulator upgrade) or if an emulator becomes unavailable. In the case of digital disk images, the significant properties to be addressed by a migration strategy usually relate to the technical interfaces identified. The functionality of technical interfaces can be documented, monitored and tested automatically, at least to certain extent. For instance, the software-side (or driver) of a technical interface to a sound card can be verified through a generic test (i.e. that sound is produced for different formats and configurations). If the test is successful then it is highly likely that any digital artefact using the same interface is also able to produce sound output. However, a general assessment of the emulator's functionality, in particular the equivalence of original and emulated CPU, is a much harder task [1]. Furthermore, computational performance features such as the rendered frame rate of graphics after processing, disk I/O performance, synchronous video and audio or even interactivity (for example, the latency between a user input event such as a mouse click and the system's visible reaction), would all need to be verified.

4.2.1 *Preservation Risks*

There are several levels of documentation possible for all three monitoring activities, depending on the choices made acquiring the software environment.

The highest level of risk for emulation exists when there is only limited knowledge about the software environment, its composition and configuration as well as hardware dependencies, i.e. technical interfaces. If, due to limited information about technical dependencies, an a-priori risk assessment is not possible, a future migration strategy may fall back to trial-and-error. Furthermore, there is a risk of an incomplete monitoring process, potentially missing obsolescence indicators. Similarly, there is a high risk of failing to rebuild the software environment if the software environment's composition and configuration is unknown. To reverse engineer an environment it is essential to have both a good knowledge of the computer system and installable software packages. Over time both these factors tend to decrease and as a consequence risk increases significantly over time. If there are resources available for technical analysis, documentation and collection software packages at acquisition, then long-term preservation risk can lowered more efficiently. This is particularly relevant for hardware dependencies and the configuration of the operating system, for instance a complete list of all the drivers installed and description of the hardware configuration used. With this information at hand an a priori assessment of technical migration risks becomes possible, as necessary drivers can be collected in advance and potential alternatives considered.

The lowest risk level is achieved when complete documentation is available about the enclosed software environment and its hardware dependencies, such that the environment can be rebuilt from scratch if necessary. In this case a preservation strategy does not solely depend on the acquired disk image (i.e. a single, "fixed" setup), as multiple strategies can be applied simultaneously. The same applies to disk images which are specifically built for preservation purposes. These images were already built within a virtualized / emulated environment, so reducing migration risk, as information on the images' content and configuration is – assuming a (semi-)automated documentation process – readily available and the installation procedures easily reproducible. The effort required for maintenance, however, may differ. This is due to different creation processes and creation goals.

Images based on documented system requirements and installation instructions replicate an existing system as closely as possible. Depending on the granularity of the available documentation, there may be slight variations and alterations to the original system specification, for example, to cope with external dependencies and/or different hardware options in the virtual machine. A further migration of these images to a new platform may require an individualized strategy, as their similarity to the original system should be maintained. In contrast, images with software installations reduced to an artwork's essential software components are more resilient to technological change, as wider variations and adaptations are acceptable, as long as the artefact can be rendered, i.e. its significant properties remain intact. In both cases, re-produced images are able to share a common technological base, e.g. share a operating system installation and configuration as well as relying on the same technical interfaces. Through a shared technological base preservation risks can be shared among similar artefacts and collecting institutions.

In general, for artefacts or software environments interacting directly with hardware, there is a higher risk of an emulation strategy failing, in particular if they rely on custom built or modified hardware components. Even if they rely on widely used hardware, not every feature (and possible quirk) of real physical hardware components may be emulated accurately. Furthermore, emulators may have bugs or behave differently compared to the original systems. In contrast, artefacts relying on operating systems to interact with emulated hardware are more likely to be successfully re-enacted using emulation, as emulator implementations usually try to cover the feature-set of popular hardware (and drivers) and/or the behaviour of a popular operating system.

4.2.2 *External Dependencies*

The management of external dependencies requires a different set of strategies, as external dependencies are technically and conceptually more diverse than hardware found in computer systems and there is usually no drop-in replacement available. Due to the diversity of external dependencies, only abstract workflows are presented.

The first, and usually most efficient strategy is internalisation of external dependencies, such that they become either a part of the digital artefact or its software environment. In general, there are two types of dependencies which can be internalized, abstract data dependencies and functional dependencies. A simple example for a data dependency is an artefact accessing data which is externally stored. An internalisation option is to copy/mirror the data and make it locally accessible, such that it becomes a part of the artefact. In general, this option is applicable for abstract data dependencies, with data being accessed or served through standard protocols, e.g. protocols directly supported by the OS. Most of the times, modifications to the object and sometime even to the software environment can be avoided. In some cases, however, changes have to be made, e.g. to point to the new location of the data (which is in particular problematic if the digital artefact used hard-coded URIs and the artefact can not be changed).

For pure functional external dependencies, e.g. a computation is made by an external machine and the artefact requires the result to perform or a software dependency requires external functionality, such as a license server, or mixed functional and data dependencies (e.g. data is served through a specific protocol, which requires additional software support such as databases), can be internalized, if the (server) machine is available and suitable for emulation. The internalized machine can then be treated as a dependent, secondary artefact, emulated and connect to the primary artefact.

A second strategy to deal with external dependencies is making technical dependencies abstract. Through abstraction the risks of failing or obsolete components to the whole setup can be reduced. The main goal is to abstract technical requirements, such that equivalent replacements can be identified and applied. For instance, a software product dependency with a dependency on a license server may be exchanged with a less problematic one, e.g. a software product providing the same technical features for a given digital file format but does not rely on a external functionality. This strategy should be included in preservation planning activities, as it may not always yield into direct useable results, but prepares the ground for future preservation actions.

Finally, emulation and simulation can be pursued, if other strategies fail or are not applicable. This strategy requires broadening the scope of emulation to include interfaces and behaviour of external components. For instance, if an artefact relies on the availability of a video-portal accessed through a dedicated web-service protocol, the protocol interface may be simulated and either translated to a newer protocol version, to retrieve content or the whole service is simulated using e.g. recorded or archived content. An example of emulated web services in the domain of research data management is provided by Miska et al [11]. A similar strategy can be applied to external hardware and hardware protocols.

5. CONCLUSION

The first step for the preservation of a software-based artwork is a technical analysis of the hardware and software setup required for its display. This analysis provides the basis for a description, which can be broken down into three conceptual layers: artefact descriptions and configuration; software environment and configuration; and hardware environment Assessing preservation options for each of these layers individually, provides a differentiated view on technological risk and potential mitigation options, and helps to make a changing technological environment more manageable.

The higher the degree to which an artefact can be abstracted from its technical environment, the more options for preservation actions remain. If an artefact can be re-built for different technical environments, its software, hardware and external dependencies may substituted (e.g. by changing the code or the artefact's setup). An artefact's software environment is of particular interest as it usually connects digital objects with hardware. If the software environment is known in its composition and configuration, the environment can, if necessary, be rebuilt in order to mitigate risk (e.g. substituting problematic dependencies). Furthermore, it becomes possible to consolidate disk images by, for example, building on a common base system consisting of operating system and a unified hardware configuration. Breaking down the technical characterization to a (common) set of technical interfaces shared among many artefacts of a similar type makes it possible to focus monitoring and technical migration work. If technical interfaces break, disk images may be migrated to a new (virtual) technical environment. Ideally, a migration path is only developed once and applied to all suitable artefacts.

In this paper we have presented an approach which, instead of looking at work-specific properties, focuses on the digital artefact's technical dependencies. If emulators were able to perfectly reproduce out-dated hardware components, this technical perspective would be sufficient – at least concerning any computational aspects of the artwork. In practice however, emulators are far from perfect in this respect, such that a manual verification of an emulated result is indispensable. For this reason, the proposed migration method relies heavily on the ability to verify the performance of a digital artwork in a new technical environment. For digital artworks any kind of automated testing of technical properties has its limitations. Software-based artworks have a second, mostly conceptual layer of significant properties, which cannot be tested in an automated way and require a specialist's assessment (for example, qualities of the artefact's behavior). Still, a structured and (partly) automated verification of an emulation's (technical) performance characteristics remains one of the most important open challenges when implementing an emulation-based preservation strategy.

Furthermore, the technical approach presented requires a wider supporting framework. Primarily, a dedicated software archive is necessary (which includes management of licenses) to help ensure that a given software environment can be rebuilt. Additionally, it is useful to maintain a testbed of typical environments and common technical interfaces to be tested on newly released emulators. In contrast to testing an artwork's work-specific significant properties, these activities, and in particular the technical infrastructure, can be shared and re-used not only for sets of similar artworks but also among different institutions.

6. ACKNOWLEDGEMENTS

This project has received support from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement number 601138 – PERICLES.

7. REFERENCES

[1] N. Amit, D. Tsafrir, A. Schuster, A. Ayoub, and E. Shlomo. Virtual cpu validation. In *Proceedings of the 25th Symposium on Operating Systems Principles*, SOSP '15, pages 311–327, New York, NY, USA, 2015. ACM.

[2] G. Brown. Developing virtual cd-rom collections: The voyager company publications. *International Journal of Digital Curation*, 7(2):3–22, 2012.

[3] M. Casad, O. Y. Rieger, and D. Alexander. Enduring access to rich media content: Understanding use and usability requirements. *D-Lib Magazine*, 21(9/10), 2015.

[4] F. Corubolo, A. Eggers, A. Hasan, M. Hedges, S. Waddington, and J. Ludwig. A pragmatic approach to significant environment information collection to support object reuse. In *iPRES 2014 proceedings*, 2014.

[5] D. Espenschied, K. Rechert, I. Valizada, D. von Suchodoletz, and N. Russler. Large-Scale Curation and Presentation of CD-ROM Art. In *iPres 2013 10th International Conference on Preservation of Digital Objects*. Biblioteca Nacional de Portugal, 2013.

[6] C. Jones. Seeing double: Emulation in theory and practice. the erl king case study. In *Electronic Media Group, Annual Meeting of the American Institute for Conservation of Historic and Artistic Works. Variable Media Network, Solomon R. Guggenheim Museum*, pages 516–526, 2004.

[7] G. Knight and M. Pennock. Data without meaning: Establishing the significant properties of digital research. *International Journal of Digital Curation*, 4(1), 2008.

[8] P. Laurenson. *Old Media, New Media? Significant Difference and the Conservation of Software Based Art.* New Collecting: Exhibiting and Audiences after New Media Art. Ashgate, 2014.

[9] T. Lurk. Virtualisation as conservation measure. In *Archiving Conference*, volume 2008, pages 221–225. Society for Imaging Science and Technology, 2008.

[10] G. S. C. Mahadev Satyanarayanan, B. Gilbert, Y. Abe, J. Harkes, D. Ryan, E. Linke, and K. Webster. One-click time travel. Technical report, Technical report, Computer Science, Carnegie Mellon University, 2015.

[11] T. Miksa, R. Mayer, and A. Rauber. Ensuring sustainability of web services dependent processes. *Int. J. Comput. Sci. Eng.*, 10(1/2):70–81, Jan. 2015.

[12] K. Rechert, I. Valizada, and D. von Suchodoletz. Future-proof preservation of complex software environments. In *Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES2012)*, pages 179–183. University of Toronto Faculty of Information, 2012.

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

[14] R. Rinehart. The straw that broke the museum's back? collecting and preserving digital media art works for the next century. SWITCH: Online Journal of New Media. http://switch.sjsu.edu/web/v6n1/articlea.htm, 2002.

[15] R. Rinehart. *Nailing down bits: Digital art and intellectual property.* Canadian Heritage Information Network (CHIN), 2006.

[16] D. S. Rosenthal. Emulation & virtualization as preservation strategies. https://mellon.org/resources/news/articles/emulation-virtualization-preservation-strategies/, 2015.

[17] J. Rothenberg. Ensuring the longevity of digital documents. *Scientific American*, 272(1), 1995.

[18] R. Russell. Virtio: Towards a de-facto standard for virtual i/o devices. *ACM SIGOPS Operating Systems Review*, 42(5):95–103, 2008.

[19] G. Wijers. To emulate or not. *Inside Installations. Theory and Practice in the Care of Complex Artworks*, pages 81–89, 2011.

A Case Study on Emulation-based Preservation in the Museum: Flusser Hypertext

Frank Padberg
Karlsruhe University of Arts and Design (HfG)
76135 Karlsruhe, Germany
fpadberg@hfg-karlsruhe.de

Philipp Tögel
Berlin University of the Arts (UdK)
10823 Berlin, Germany
variantology@digital.udk-berlin.de

Daniel Irrgang
Berlin University of the Arts (UdK)
10823 Berlin, Germany
irrgang@medienhaus.udk-berlin.de

Martin Häberle
ZKM Center for Art and Media Karlsruhe
76135 Karlsruhe, Germany
martin.haeberle@zkm.de

ABSTRACT

We use emulation to preserve a complex digital artifact in the museum. We describe all stages of the preservation process and discuss the technical and curatorial problems that we encountered. The exhibition setting defines additional requirements for an emulation. Our findings and lessons learned are instructive for emulation researchers and museum practitioners. The preserved artifact now is on display in real exhibitions.

Keywords

Emulation; Digital Preservation; Digital Art; Museum Exhibition; Virtual Disk; Hypertext Systems; HyperCard; Vilém Flusser; Media Theory

1. INTRODUCTION

We report on a successful effort to preserve a complex digital artifact of the early 90s and present it in a museum exhibition using emulation. We describe the instructive difficulties that we encountered at all stages of the preservation process, ranging from the preparation of the digital artifact to the operation of the emulation in the exhibition. We also comment on the technical and curatorial decisions that we made.

We suppose that our observations and findings are quite typical for an emulation-based approach, and that the lessons learned in our case study will prove useful for emulation researchers and museum practitioners.

The digital artifact that we aimed to preserve is the so-called "Flusser Hypertext" from 1992. This work not only is part of the digital heritage of the philosopher and media theorist Vilém Flusser, but also an important document of the early history of hypertext systems. The Flusser Hypertext project started at a time when the Web as we know it today was still in a very early phase of its development. Few guidelines had been established at that time for how to structure and lay out hypertext content, and a lot of experimentation was going on.

The Flusser Hypertext lent itself to an emulation approach since it originally executed on a standard computer of its time for which an emulator is available "off the shelf," and it had no special hardware or interface requirements. An earlier attempt to display the Flusser Hypertext at the art festival "Transmediale 2010" using a vintage computer had to be suspended, because the old hardware turned out to malfunction too frequently when operating over days. Hence, this time we opted for an emulation of the vintage environment on modern, reliable hardware.

The preserved Flusser artifact was actually presented in public as part of the retrospective exhibition "Without Firm Ground – Vilém Flusser and the Arts," shown from August to October 2015 at the ZKM Center for Art and Media in Karlsruhe, and from November 2015 to January 2016 at the Academy of Arts in Berlin. The exhibition is currently on display until May 2016 at "The West," a gallery and art museum in Den Haag, Netherlands.

The emulation proved stable during its many weeks of operation in the exhibition. Concerning the preservation process, main summary findings are:

- The lion's share of the effort went into the analysis of the run time environment required by the artifact, and the preparation of the virtual disk.
- The preservation required deep technical knowledge and definitive curatorial judgments at all stages.
- The exhibition setting posed additional challenges for the emulation.
- True to original hardware was a valuable, sometimes indispensable tool in the preservation process.

This paper presents a single-case study, but our findings are supported by our previous and ongoing experiences with preparing multimedia and digital art objects for an emulation-based presentation in the museum.

2. RELATED WORK

Emulation has been studied and discussed for 20 years [1] as a preservation technique for digital objects, and it already was the subject of substantial research efforts (see, f.e., [2][3][4]). Yet, emulation seems to have been applied mainly to multimedia in libraries [5][6] and computer games [7][8], but not much in software-based art. There also seems to be a technical research focus on automated emulation frameworks [9][10][11][12][13].

In the art museum, emulation seems to have been employed only rarely as yet. As a result, there is a lack of concrete observations from real exhibitions, and the amount of practical advice available to museum practitioners is very limited.

Rinehart and Ippolito [14] report on a symposium and exhibition in 2004 centered about various uses of emulation in games and art. Among other things, three pieces of software-based art were actually emulated one-one. The emulations were compared against the originals, and the impact of emulation on the appearance of the art works was discussed.

Kaufmann [15] in 2011 describes the hardware emulation of a home computer cartridge that contained a rare, but important software art object; the emulated cartridge was used in the subsequent exhibition. Padberg [16] in 2014 shows a software emulation of the same art object in contrasting juxtaposition with a true-to-original version of the work. Falcao e.a. [17] in 2014 briefly sketch an in-house, exploratory trial to preserve two digital objects from their art collection using virtualization.

The online museum rhizome.org [18] presents a growing number of digital objects using emulation. Their emulations must run over the Web and, hence, are restricted to objects that have minimal interface requirements, requiring just a display, keyboard, mouse, and sound. Clearly, presenting online is much different from the typical museum exhibition in real spaces, and this fact has a strong impact on the look-and-feel of the emulated art object. Similarly, Espenschied e.a. [19] in 2013 present a case study in which six selected software-based artifacts of the "Transmediale" art collection on CD were emulated online using their web service-based emulation framework.

Lurk, e.a. [20] in 2012 discuss requirements for the emulation of digital heritage objects, including digital art. Lorrain [21] very briefly reports on a failed emulation of a software-based art work in 2010. Besser [22] in 2001 discusses the effect of exchanging vintage I/O devices against modern ones on the appearance of digital art objects.

3. ARTIFACT

The Flusser Hypertext is based on the 1989 lecture "Schreiben für Publizieren" [Writing for Publishing] by the Czech philosopher and media theorist, Vilém Flusser (1920-1991) ([23] p.510). The lecture was given at the Institute for Technology Assessment and Systems Analysis (ITAS) of the Karlsruhe Nuclear Research Center (today part of the Karlsruhe Institute of Technology).

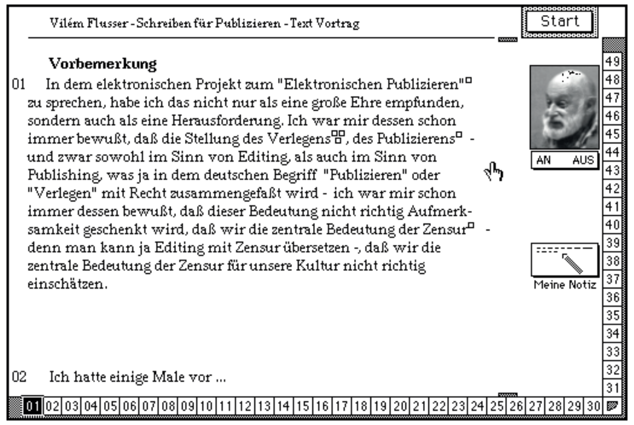
The Flusser Hypertext was developed as part of the ITAS research project "Elektronisches Buch" [Electronic Book]. The project goal was to conduct research on the conceptual and technological possibilities of an "innovative electronic presentation form for results of scientific projects" [1] [24]. The ITAS team envisioned to develop a "multimedia study system" that would use Flusser's lecture (which was accessible as both audio recording and text) as a starting point, and to expand it with additional information and interactive elements.

To transform Flusser's lecture into a multimodal hypertext means to take Flusser's theories serious: transferring the spoken word into the electronic text domain, while enriching it with other forms of media (sound, images) and interactive elements [25]. Flusser himself described multimodal electronic media as transitional phenomena, preparing the dawn of an "universe of technical images" [26]. By building upon his earlier language philosophy and communication theory, Vilém Flusser's media theory of the 1970s and 1980s can be read as an analysis of the coming informatized society, claiming (similar to Marshall McLuhan) the end of writing as the dominant discursive form. According to Flusser, written text and the "linear" structure of discourses will vanish and soon be replaced by what he called "synthetic images": visualizations of concepts that need not be transcoded into letters but can be "calculated" as computer generated images.

The Flusser Hypertext was developed by the team of the ITAS (Knud Böhle, Ulrich Riehm, and Bernd Wingert) and a team of freelance programmers. Vilém Flusser was not directly involved in the development. However, he did supply further information on and explanations of references that he made during his lecture. Bernd Wingert demonstrated an early version of the Hypertext in May 1991 at Flusser's home in Robion, France (cf. [24] p.209 and [27] p.109). According to Bernd Wingert[2], Vilém Flusser was obviously honored to see his words being adopted by the technological apparatuses which he had been theorizing about for years.

The Flusser Hypertext was never finalized or officially published. The version discussed in this paper reflects the last state of the project work from 1992. Live versions of the Hypertext prototype were demonstrated and discussed at six different public venues between 1990 and 1993 ([28] cf. [29]; [30] cf. [24] p.161; [31]; [32]; [33] cf. [24] p.194; [34][35]).

The Hypertext was programmed using Apple's "HyperCard" system, an early, general-purpose authoring system that supports multimedia content and allows for programmed layouts. The backbone of the work is the transcript of Flusser's lecture. The transcript is organized as a so-called stack of cards, that is, the text is subdivided into slices that fit onto a single, fixed-size HyperCard screen. The cards are linked by number, and the user can browse through the cards by clicking on the numbers at the bottom and right edge of each card, see Figure 1.


Vilém Flusser - Schreiben für Publizieren - Text Vortrag

Vorbemerkung

01 In dem elektronischen Projekt zum "Elektronischen Publizieren"□ zu sprechen, habe ich das nicht nur als eine große Ehre empfunden, sondern auch als eine Herausforderung. Ich war mir dessen schon immer bewußt, daß die Stellung des Verlegens□□, des Publizierens□ - und zwar sowohl im Sinn von Editing, als auch im Sinn von Publishing, was ja in dem deutschen Begriff "Publizieren" oder "Verlegen" mit Recht zusammengefaßt wird - ich war mir schon immer dessen bewußt, daß dieser Bedeutung nicht richtig Aufmerksamkeit geschenkt wird, daß wir die zentrale Bedeutung der Zensur□ - denn man kann ja Editing mit Zensur übersetzen -, daß wir die zentrale Bedeutung der Zensur für unsere Kultur nicht richtig einschätzen.

02 Ich hatte einige Male vor ...

Figure 1. Sample text card

When audio gets activated by the user, the text cards are underlaid with the live recording of Flusser's lecture: The fragment of the recording corresponding to the current card is played, and at the end of the card the system automatically moves to the next card.

The transcript is augmented by hypertext links that open separate cards containing bibliographical references, short articles, or annotations explaining particular topics and names mentioned in the lecture, similar to today's web links. The articles were provided by experts in the field, including Flusser himself. This supplementary material (about 450 cards) far exceeds the main text in size (49 cards). Additionally, users have the option to add their own notes to any card: The note pad opens; any text will be saved automatically and can be edited later.

Overall, the Flusser Hypertext is clearly structured. There is no poly-hierarchical network of links but a structure which could be described as "horizontal and vertical" ([24] p.187-188): The cards in horizontal order present the lecture, the cards in vertical order contain the supplements; see Figure 2. Each vertical "string of cards" is separated from other vertical strings – there are no links connecting articles or annotations from different text cards.

[1] all quotes were translated from German by Daniel Irrgang

[2] personal conversation with Daniel Irrgang and Philipp Tögel

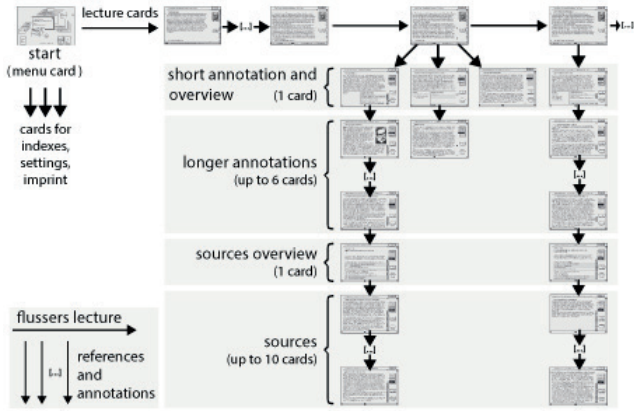

lecture cards
start
(menu card)
cards for indexes, settings, imprint
short annotation and overview (1 card)
longer annotations (up to 6 cards)
sources overview (1 card)
sources (up to 10 cards)
flussers lecture
references and annotations
[...]

Figure 2. Link structure of the Hypertext

As Peter Wiechens concludes in a case study [36] on the Flusser Hypertext: The Hypertext does not put an end to the lecture's linear structure but rather adds a dimension of simultaneity due to its multimodal features (text, image, sound). Simultaneity and multi-modality, as well as the "instantaneous availability" of further information (the annotation and reference cards) are pointed out as the Flusser Hypertext's pivotal aspects in the ITAS research report [24]. One might add to this list the possibility to add own notes to each card. For Vilém Flusser, this kind of productive engagement – the reader being put in the position of an author – was certainly a striking feature that he emphasized in his reflections on the Flusser Hypertext [37]: "The original lecture level falls into oblivion, overwhelmed by comments and counterarguments." [1]

4. PREPARATION

For an emulation, a so-called virtual hard disk is required as input. This is a file (on the host computer) that has the same internal structure as the hard disk of the vintage, to-be emulated system. The virtual disk also contains a copy of the vintage operating system, which will be booted automatically by the emulator.

There are basically two ways to obtain a virtual disk: either by using a tool that creates a file with such a structure and allows for installing the vintage operating system on this file; or, by taking a one-one file image of an existing vintage hard disk. In our case study, we started with a hard disk image.

The digital artifact can be put in two places: either it is placed in a separate file having a standard format such as an ISO file that serves as additional input for the emulator; or, the artifact's files get copied to the virtual disk. In our case, the hard disk image already contained the Flusser Hypertext files.

Although we started from a one-one image of a real hard disk, we encountered technical problems when preparing a virtual disk for the emulation. We had to take additional measures to adjust the run time environment of the Flusser Hypertext on the virtual disk. Exceptionally for an emulation, we even patched the code of the artifact. We also had to de-activate certain extensions of the vintage operating system so that the emulator booted the virtual disk. This is a common problem, since many emulators do not support all features of a vintage system.

4.1 Creating the Initial Virtual Disk

The Flusser Hypertext is a research prototype that never advanced into a production-quality release. Only a single copy of its code and data files still existed, stored as a snapshot of the last development version[3] on the physical hard disk of a vintage Macintosh Performa 630 computer. The vintage Mac was donated to the Flusser Archive in Berlin in 2007.

As the first step in the preparation process, which began in June 2013 at the Flusser Archive in Berlin, we created a one-one image of the vintage hard disk and stored it in a file. Detaching the hard disk from the vintage computer and attaching it to a standard PC usually is the best option, see Figure 3.

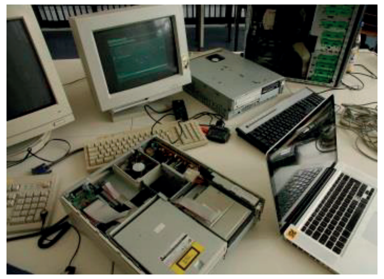


Figure 3. Cloning the hard disk[4]

The image file contains both the vintage operating system and the Flusser files and, basically, can serve as the virtual disk for the emulation. When trying to boot this first image file as-is with the emulator (see section 5), the emulation crashes on loading the system extensions. We solved this problem [38] by creating another disk containing a fresh installation of the vintage operating system, from which we could access the first disk image and disable the system extensions in question ("Video Startup" and "A/ROSE").

4.2 Problems in the Run Time Environment

Tests on the Performa computer uncovered a reproducible error that leads to a crash of the HyperCard application: When invoking the audio playback on certain text cards (33-42), the application displays an error message "Unerwarteter Fehler 5454" [unexpected error 5454], forcing the user to quit the application,

see Figure 4. After re-launching the application, the buttons for selecting the individual cards and for playing audio do not react until the user either hits the "stop audio" button or navigates to the Hypertext's reference layers, and back again.

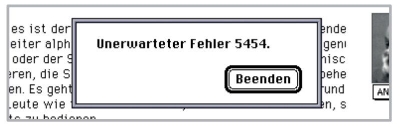

Unerwarteter Fehler 5454.

Figure 4. Unexpected error

While the previous error already occurred on the original computer, the following problem could only be observed when copying and running the Flusser Hypertext files on other machines (even of the same model series), or when running them in the emulator: The Hypertext executes, but fails to output any sound at all. The audio function of the HyperCard framework as such was tested, and it worked on all systems. All

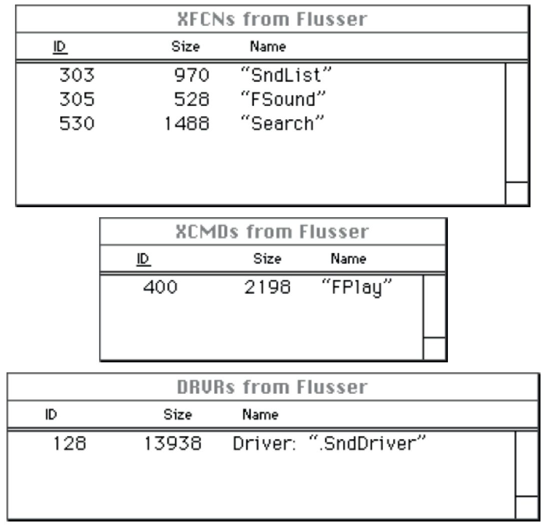
[3] Incomplete supplementary cards of the development version had subsequently been filled with content by Bernd Wingert (between 1994 and 2007). No cards were added, and the structure was left unchanged.

[4] image taken from the Vilém Flusser Archive collection (© Vilém Flusser Archive)

attempts to solve this problem by altering the run time environment failed. The original machine seemed to remain the only system capable of playing the Hypertext's audio content.

Online sources indicate that error 5454 relates to corrupted HyperCard files. We traced the defect back to a particular file ("Ton 8") that holds the audio data of the affected text cards. This file produced the same error when opened in HyperCard directly.

We found a way to inspect the Hypertext's source code (written in HyperCard's scripting language HyperTalk). We traced function calls that relate to audio, and found some functions ("FSound", "Search", and "Fplay") that were neither defined in the code, nor part of the HyperTalk language. Instead, these functions are contained in precompiled libraries which get stored along with the Flusser files, see Figure 5 . We also found a special driver library (".SndDriver“) that might possibly be involved in the error.



XFCNs from Flusser

ID	Size	Name
303	970	"SndList"
305	528	"FSound"
530	1488	"Search"

XCMDs from Flusser

ID	Size	Name
400	2198	"FPlay"

DRVRs from Flusser

ID	Size	Name
128	13938	Driver: ".SndDriver"

Figure 5. Precompiled libraries and drivers

Apparently, we had encountered a highly intricate problem with the specific run time environment in the form of these libraries.

4.3 Patching the Artifact

A reverse engineering of the precompiled libraries was beyond the scope of the project, both technically and with respect to time and effort. Thus, we made the decision to avoid all calls to unknown functions in the Hypertext code, whilst keeping its functionality unaltered.

To eliminate the need for custom audio-related code, we ported [38] the Flusser Hypertext from HyperCard 2.0 to version 2.3, whose "ADDmotionII" standard library provided the desired audio functionality. By replacing the artifact's special audio functions with the standard library functions, we succeeded in enabling audio playback in the emulation as well as on our vintage Mac computers.

Having access to the artifact's code also enabled us to solve the error 5454 issue: In HyperCard, we built a replacement file for the seemingly corrupted "Ton 8" file from scratch, using the original audio resources, which we had salvaged from the "Ton 8“ file.

As a result of the whole process we obtained a virtual disk that boots with the emulator and is fully functional. The patched artifact shows no unexpected error messages or even crashes. Otherwise, it appears almost unaltered to the user. In particular, the patches are not obvious to the user as an amendment to the artifact. The only minor exception is a short interruption of the playback of the audio recording (see section 3) in the middle of each text card, when HyperCard loads the second half of the audio data for the card. This deviation results from the patches that we applied to enable audio playback, and could not be avoided.

The decision to accept such slight deviations was reached by the exhibition curators who were aware of the conservational implications of the technical patches. Yet, a guiding theme of the exhibition was to reveal the continuing effect of Flusser's philosophical writings on current phenomena in arts and media. An early, striking example for such a crosslink was the Flusser Hypertext, an innovative "multimedia study system" that reflected Flusser's ideas (see section 3). Hence, the curators favored a smooth user experience and the stability of the emulation during the exhibition over absolute fidelity to the inherited object at the code level.

The patches are marked by comments in the code, but cannot be reversed at the push of a button. Certainly, we saved different intermediate versions of the code during the patching process, including a completely unaltered disk image which can serve as the starting point for more traditional conservation approaches.

5. EMULATION

The vintage Macintosh computer that contained the snapshot of the Flusser Hypertext features a Motorola 68040 CPU running System 7.1.2 as its operating system. This type of computer can be emulated using the "BasiliskII" emulator that is freely available for Windows, Linux, and OSX hosts. The Flusser Hypertext poses no special hardware or interface requirements to the emulator.

At the beginning of the preservation project, we had little experience with this particular emulator. We asked colleagues from the University of Freiburg for support, who kindly helped by setting up the initial emulation.

5.1 Emulation during Preparation

Once we had managed to produce a virtual disk that was bootable with the emulator (see subsection 4.1), we made extensive use of emulation during the preparation phase. We found this to be a convenient approach since our artifact required an unusually long trial and error-phase of changes to its run time environment and even code (see subsection 4.2). If some change failed, it was much easier to return to the last working copy of the virtual disk under emulation than to restore the previous state of the real hard disk in the vintage computer.

We used a stand-alone setup of the BasiliskII emulator during preparation. A stand-alone setup must be manually configured, but provides fast emulator start-ups, direct access to the emulator's configuration parameters, and easy restoration of any input file that did not work as desired or got damaged in a trial run, by simply overwriting the file with a backup copy.

Before making any changes permanent, we tested on our vintage computers whether the changes had any unwanted impact on the behavior of the artifact.

5.2 Configuring the Emulation

For the actual exhibition, the BasiliskII emulator was packaged into a stand-alone, bootable-from-stick version of the Freiburg emulation framework [12].

The bootable stick comprises of a special boot loader and two partitions. One partition holds the Linux host system, the emulation framework, and the emulator. The other partition holds the virtual disk file with the patched Flusser Hypertext files, a Mac Quadra 630 ROM image, and the configuration data for the emulator.

Using this pre-fabricated stick had practical advantages for us: The Linux host system boots automatically when power is turned on at the mini PC that served as the host computer in the exhibition. The Linux system also was configured to automatically start the emulator at boot time with (a fresh copy of) the virtual disk. In addition, some keyboard shortcuts were disabled to prevent users from gaining access to the underlying Linux host system.

The screen resolution in the emulator was set to 640x480 pixels, a resolution which is typical for Mac computers of the time. The color depth was set to maximum. Otherwise, standard values were used for the configuration parameters of the emulator (see subsection 5.5).

5.3 Sound Problem

Despite its benefits, the automated framework approach turned out to have significant drawbacks: Since the emulator was completely encapsulated within the framework, which booted immediately to the emulation when turning power on, there was no direct access to the configuration of the emulator, nor to the virtual disk.

This caused problems right before the exhibition started, when we discovered that the sound was missing in the emulation.

The sound must be activated at all levels of the emulation: in the host system, in the configuration of the emulator, in the options (if any) of the artifact, in the emulated Mac system, and, finally, at the speakers. The problems were resolved in Freiburg.

For future exhibitions, we are seriously considering to use a stand-alone emulator in the exhibition instead of a packaged emulator, providing for more direct, in-house control over the emulator configuration.

5.4 Peripheral Devices

The screen resolution for the emulation is 640x480 pixels. Presenting such a low resolution on a modern display results in a small area on screen; alternatively, scaling this up to the size of a modern display results in a blurred or even distorted picture. In addition, the curatorial goal was to preserve as much of the original look-and-feel of the Flusser Hypertext as possible in the emulation.

Hence, the curatorial decision was made to use original, resp., true to original peripheral devices in the exhibition.

We used a 15 inch vintage Apple multiscan color display. This particular model is slightly more recent (1994-96) than the Flusser artifact, but it comes with a VGA port instead of the more typical Apple DB-15 port, which made it easy to connect the display to the mini PC using a standard VGA cable. No special adapter was needed, as opposed to 14 inch Apple displays.

We also used a vintage Apple extended keyboard and Apple mouse. To connect them to the mini PC, an Apple Desktop Bus (ADB) to USB adapter was required. Such adapters are still available over the Internet for a reasonable price.

For the sound output, we used a pair of external vintage speakers placed next to the display and connected with standard audio cables to the mini PC.

Only the display, keyboard, mouse, and speakers were placed on top of the table used in the exhibition — the mini PC was hidden underneath the table (see Figure 6) and was not easily visible to the visitors. The whole arrangement looked quite authentic; as if it came straight from the 90s, see Figure 7.

Figure 6. Mini PC under the table[5]

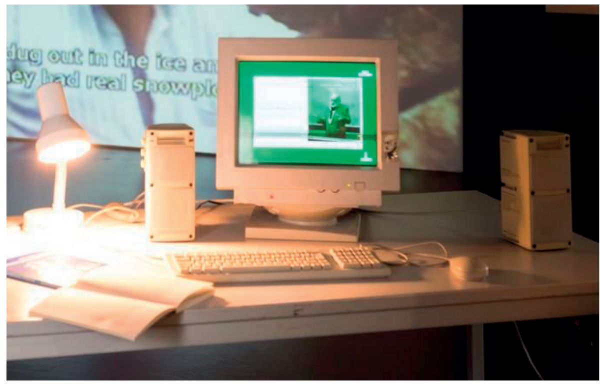
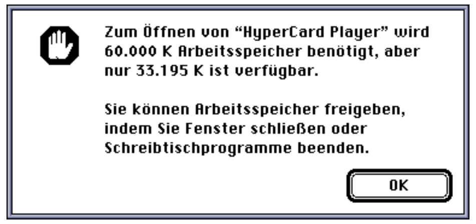


Figure 7. Emulation in the exhibition[6]

5.5 Memory Configuration

In an emulation, the main memory (RAM) required by the vintage operating system for its execution gets emulated; that is, it is provided by the emulator program through software means. The amount of main memory to be emulated must be configured as a parameter of the emulator.

When setting this parameter to a common value of 32 MB (the Performa 630 has a physical maximum of 36 MB), our virtual disk booted under emulation, but the Hypertext failed to start, complaining about not having enough memory, see Figure 8.



Zum Öffnen von "HyperCard Player" wird 60.000 K Arbeitsspeicher benötigt, aber nur 33.195 K ist verfügbar.

Sie können Arbeitsspeicher freigeben, indem Sie Fenster schließen oder Schreibtischprogramme beenden.

Figure 8. Not enough memory

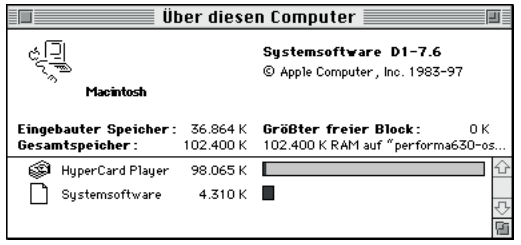
When an application requires more memory than is available as physical RAM on a vintage Mac, the administrator can reserve part of the hard disk as so-called virtual memory, which then is added automatically by the Mac operating system to the total memory available to applications.

In the BasiliskII emulator, support for virtual memory is not implemented, though: When activating virtual memory in the emulated Mac, the emulator crashes. The quick solution is to simply set the amount of *emulated* RAM to the desired total value. For the Flusser Hypertext, setting the emulator's memory parameter to 128 MB works; we used this value in the exhibition.

At a later occasion, though, we observed that only a small fraction of the reserved memory actually gets consumed by the HyperCard application that processes the Flusser Hypertext, see Figure 9.

[5] image © Martin Häberle

[6] image taken from https://www.itas.kit.edu/2015_032.php (© Karlsruhe Institute of Technology)


Über diesen Computer

Systemsoftware D1-7.6
© Apple Computer, Inc. 1983-97

Macintosh

Eingebauter Speicher: 36.864 K Größter freier Block: 0 K
Gesamtspeicher: 102.400 K 102.400 K RAM auf "performa630-os...

HyperCard Player 98.065 K
Systemsoftware 4.310 K

Figure 9. Actual memory consumption

On a vintage Mac system, the amount of memory that a program claims at start-up is preset and can be viewed in its "Information" window. For the HyperCard program, this value was set to 60,000 KB, which is far more than is actually needed for processing the Flusser files. Changing this value to more moderate 20,000 KB is completely sufficient and eliminates the need to use virtual memory on the vintage Macs, resp., to use an exceedingly high value for the emulator's memory parameter.

This issue demonstrates that settings in the emulated operating system can feed back into the configuration of the emulator, and vice versa.

6. OPERATION

The emulation operated continuously for more than 4 months in the exhibition. During this period, we made a number of instructive observations concerning the proper setup and operation of emulations in a museum and exhibition setting. In some cases, we learned that we should better do certain things differently in future emulation-based exhibitions.

We also extended our comparison of the look-and-feel of the Flusser Hypertext under emulation against its original appearance when executing on restored, vintage Mac computers. This allowed us to better assess the quality of the emulation, and it provided additional observations about the run time environment required by the artifact.

6.1 Deleting the Object

After booting-up the host system, the emulation automatically starts (see subsection 5.1). The emulator presents the desktop of the vintage Mac system to the user, including an icon in the center for starting the Flusser Hypertext (see Figure 10), waiting to get double-clicked. The desktop also offers two supplementary videos that explain the usage of the Flusser Hypertext; the videos were produced in the 90s along with the Flusser project.

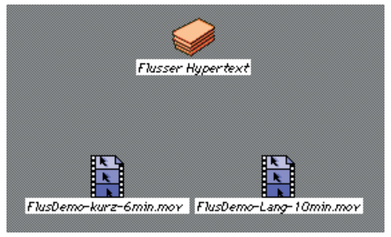


Figure 10. Desktop icons for the Hypertext and videos

Clearly, this is *not* a fail-safe setup for a public exhibition. Instead of starting the digital artifact, the user can access various Mac system functions using the menu bar at the top edge of the desktop. The user can even *delete* the object by dragging it into the waste basket, see Figure 11. This will leave the emulation in an unusable state for the next visitor, who will probably be confused and turn away from the object without telling anybody.

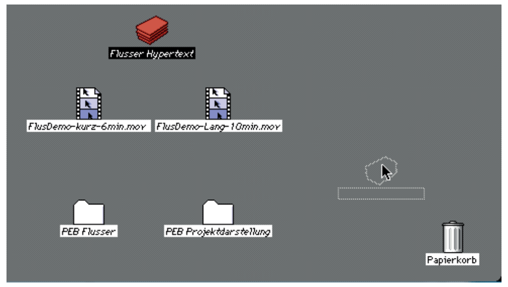


Figure 11. Draging the artifact into the waste basket

When the object has been deleted, a functional state can only be recovered by re-booting the emulation. This problem actually occurred during the Karlsruhe exhibition. In the sequel, the technical staff kept an eye on the Hypertext exhibit to take corrective action if necessary. For the exhibition in Berlin, the supervisory staff was asked to check the state of the emulation several times a day; if something was wrong, they re-booted the whole system.

Our approach was motivated in part by the desire to make the videos accessible to the visitors, but we might better have presented the videos on a separate computer and display.

For future exhibitions, a more elaborate and fail-safe approach is needed. In a museum setting, not only the emulator should start automatically, but even more so the digital object itself. Users should not be allowed to exit the running object. If the exit function cannot be de-activated within the object itself, the object must get restarted automatically and immediately by the system whenever a user chooses to quit.

This automatic restart-feature for the object must be implemented at the level of the emulated operating system, combining autostart features of the vintage system with a special "watchdog" program that runs in the background to supervise the object in question: If the object (more precisely, its application process) stops running for some reason, the "watchdog" restarts the object.

For some digital objects, it may also be appropriate to restart the object whenever it has been inactive for some time, because visitors typically just leave one exhibit and move to the next. An automatic reset will always present a tidy object to the next visitor. Inactivity can be detected using a watchdog programmed with a time-out.

6.2 Exit

After the emulation has started, the desktop of the vintage Mac system is presented to the user. Similar to the problem of deleting the digital object as described in the previous subsection, a user can also deliberately shut down the emulated Mac system using the menu bar on the Mac desktop. This occasionally occurred in the Karlsruhe exhibition. When the emulated Mac shuts down, the emulator program itself exits automatically, and control returns to the underlying host system. The emulation framework then presents a screen that asks the user to restart the emulation.

Again, this is *not* a fail-safe setup for a public exhibition. There should be no way for museum visitors to exit the emulation or shut down the emulated system. If the shutdown functionality cannot be de-activated at the level of the vintage system, the emulation must restart immediately without manual intervention. This can be achieved using standard start-up and process control features of the underlying host operating system.

6.3 Pixel Errors

In the emulation, some spots occur in the small Flusser photograph that is placed at the right hand side of the text cards (see Figure 12). At first, we thought that these spots were pixel errors introduced somehow by the emulation. Yet, a quick comparison revealed that the spots are also visible when running the patched artifact on a vintage Mac; that is, the pixel errors were nothing but false positives.

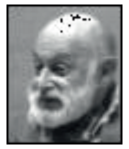


Figure 12. Pixel errors

This issue serves as an example that original hardware not only is an indispensable tool to identify deviations of the emulation from the original, but, contrary, also helps to confirm that the emulation actually *conforms* to the original appearance.

6.4 Missing Fonts

A problem that we often encounter in emulations of digital objects is missing fonts. For the Flusser Hypertext, this initially seemed to be no issue, since we used a hard disk image of the vintage computer as the basis for the virtual disk in the emulation, and that image contained all necessary add-ons, including fonts. But when installing the Flusser data and code files on our restored vintage Macs, which carried a fresh system installation, the missing font problem popped up: As the vintage operating system substituted a standard font for some missing fonts automatically, the text on each card was incomplete and the layout looked somewhat distorted.

The problem would also have occurred if we would have taken the (quite common) approach of using two virtual hard disks for the emulation instead of one, separating the operating system to be emulated from the files of the digital object.

Under the Mac operating system, required fonts must be copied to the proper system folder. Our disk image contained several dozens of fonts that had been added to the development system after installation; hence, determining the exact set of required fonts was quite time-consuming. The Flusser Hypertext actually requires 3 non-standard fonts.

This example shows that original hardware is often helpful to identify additional, hidden dependencies of the digital artifact on its run time environment.

6.5 Pace of Operation

We conducted a few experiments to evaluate whether the emulation shows any noticeable difference in its pace of operation as compared to the Flusser Hypertext running on a vintage Mac computer. The emulator executed on a standard laptop (dual-core at 1.9 GHz, 4 GB of memory, Windows 7); the vintage Mac was a restored Performa 630 (68040 at 20 MHz, 36 MB of RAM, no virtual memory, Mac system 7.6).

The Flusser Hypertext is largely static in nature, changing screen only when the user interacts with the program; typically, when moving to the next card by clicking on a number at the bottom or right edge of the current card. The only "animated" behavior can be observed when the recording of the lecture is played. In this mode, the system automatically moves from card to card, in sync with the recording, see section 3.

There is no noticeable speed difference when listening to any individual card. Overall, the emulation is slightly faster than the original Mac, lying ahead by 1.5 text cards at the end of the whole lecture, which consists of 49 text cards. On the vintage Mac, there is a slightly longer delay when the system moves from card to card. The emulator running on modern hardware seems to be faster when it comes to loading the next data into the HyperCard program.

The total gap is much larger when comparing the Performa 630 (no virtual memory) against another vintage Mac (Quadra 650, 68040 at 33 MHz, 36 MB of RAM, Mac system 7.6) that has virtual memory enabled. Although the Quadra is the faster computer, the virtual memory mechanism slows down the loading of data significantly. The Performa with no virtual memory is ahead by 1 card already after about 1/3 of the lecture.

The experiments show that the audio sequences are replayed faithfully in the emulation, which is a key factor for the authenticity of any emulation. In addition, we learned that variation in the vintage hardware or configuration can have a much larger impact on the appearance of a digital object than the emulation. Such measurements are impossible to conduct without having original hardware at hand.

6.6 Hanging Print Function

Except for the text cards containing the transcript of Flusser's lecture, all cards in the Flusser Hypertext can be printed out using an icon in the lower right corner of each card, see Figure 13. This includes the cards containing supplementary material for Flusser's lecture, such as explanatory articles or bibliographies, and the personal note cards created by the user (see section 3).

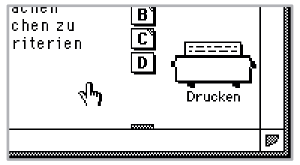


Figure 13. Print function

In the exhibition, no printer was connected to the host computer, and the emulation was not configured to accept and handle print requests. Nonetheless, it was possible for a visitor to click on the printer icon and create print jobs. If done repeatedly, a long print queue emerged that blocked the user interface – the program became irresponsive and seemed to "hang."

The Flusser Hypertext does not offer an option for deactivating its print function similar to deactivating its audio function in the settings. Hence, either a real printer must be added and the emulation configured accordingly, or, a non-blocking "mock printer" must be installed at the level of the emulated vintage Mac system. The problem was discovered only late in the Karlsruhe exhibition. For the Berlin exhibition, we left the virtual disk as is and relied on our supervisory staff to handle any problems.

This issue highlights the fact that digital objects can include features that must be explicitly handled at the technical level in an emulation. For objects with a complex internal structure, such features need not be as obvious as the print function, but can be buried rather deep inside the object. F.e., in other objects under preparation we encountered hidden links to the internet. Such features might even be undocumented, especially in digital art that often includes elements of surprise which the user is expected to discover when interactively exploring the piece of art.

7. CONCLUSIONS

In this paper, we presented a real case study that illustrates the whole process of preserving a digital object in the museum by means of emulation. The required steps range from creating the virtual disk to operating the emulation in a public exhibition.

We encountered a number of instructive technical and curatorial problems that resulted in a number of specific lessons that we learned, some typical for any emulation-based preservation, others typical for the exhibition setting that we worked in.

(L1) *Preparing an artifact for the emulation demands a close analysis of the run time environment needed by the artifact*.

This includes the identification of special drivers, non-standard support files, external interfaces, and special system settings (cf. subsections 4.2, 5.3, 6.4, 6.6).

The ties of a digital object into its run time environment can be very subtle; one missing detail can lead to a strange behavior or failure. In our experience, the required run time environment often is not documented in sufficient detail even for artifacts published on distributable media, turning the analysis into a trial-and-error process.

(L2) *Emulators typically do not support every feature of the vintage system*.

Under emulation, drivers may fail to work or crash the emulator, system options may crash the emulator when activated, or external interfaces may be unavailable. This creates problems when trying to install and/or execute an object with its run time environment under emulation (4.1, 4.2, 5.5).

In practice, it is not always possible to pin down and fix the root-cause of a vintage system-level problem. It may be necessary to circumvent the problem by de-activating system features, exchanging certain drivers, or moving to another version of the vintage application program and operating system. Patching the digital object (as we did) should be a last resort.

(L3) *Emulating a whole system-image can be a non-trivial task*.

A full image might reflect an intermediate version of a digital object (as in our case), or conserve a complete heritage work environment. Assuming a non-networked system, the image will contain all required run-time components; yet, some components may not work under emulation (4.1, 4.2).

The complexity of a full image of a "living" system exacerbates the difficulties of fixing inconsistencies in the run time environment under emulation. When the artifact to be emulated is not finalized, tested software, it may contain bugs, and its documentation will likely be fragmentary or missing, adding to the problem.

(L4) *In the museum, curatorial judgments provide the direction for the technical setup of the emulation*.

Curatorial judgments refer to the choice of peripheral devices, the acceptance or rejection of deviations of the emulation from the original, the user interface offered, and the acceptance of any changes to the artifact (5.4, 6.5, 6.1, 4.3).

An emulation need not necessarily be perfect. Augmenting the emulation with true to original peripheral devices can provide a way to preserve the essentials of the original look-and-feel.

In our case, even patches to the artifact seemed admissible from a curatorial perspective, given two facts: The Flusser Hypertext was not a final, production-quality release, but an advanced prototype; the computer on which it was stored was not Flusser's own personal computer, hence, it made little sense to try and preserve something like Flusser's "digital working environment."

(L5) *In a museum exhibition, the emulation must be specifically safe-guarded*.

In a public exhibition, the continuous operation of the emulation must be guaranteed. This poses technical challenges for the setup of the emulation, in excess of configuring the emulator. The artifact itself must be protected against deletion, and the emulation must be sealed to prevent any unintended usage of the emulator, vintage system, or underlying host system (6.1, 6.2, 6.6).

Visitors should best be prevented from quitting the running artifact at all. This requires elaborate technical measures at the vintage and host system level.

(L6) *Preparing an emulation requires substantial technical knowledge and skills*.

Knowing the configuration options of the emulator certainly is a prerequisite to tailor the emulation to the object (5.2, 5.5). In addition, in the preparation phase frequent trial runs of the emulation are typical, until a working overall setup is found (5.1, 5.3, 5.5). To achieve short cycle times, a stand-alone emulator is better suited than a framework, which encapsulates the emulator and input files into additional software layers and special workflows. Yet, such a stand-alone setup must be manually configured.

Knowledge of the emulator is not sufficient to prepare the one central input for the emulation: the virtual disk. The virtual disk contains the vintage run time environment of the object to be emulated. An understanding of and practical experience with the vintage operating system, its hardware and application programs is indispensable for solving any problems of incompatibility of the digital object with the emulator, by customizing the virtual disk (cf. L2). This also applies to safe-guarding the emulation in an exhibition (cf. L5). In addition, settings in the vintage system can feed back into the configuration of the emulator, and vice versa (5.5, 5.3).

(L7) *The preparation of a virtual disk for the emulation consumes the lion's share of the total effort*.

The virtual disk contains the run time environment for the emulated artifact, compensates any shortcomings of the emulator, and reflects the special technical measures taken for an exhibition setup. The more complex the digital object, the larger the preparation effort. Even for more average objects of digital art, which do not require analyzing and patching the artifact's code as in our case, we found that the effort for preparing the virtual disk (including the time-consuming trial runs of the emulation during preparation) typically is substantial and far exceeds the effort for installing and configuring the emulator. This fact seems to get severely underestimated in the literature.

(L8) *Original hardware is a valuable, sometimes indispensable tool in the preservation process*.

Comparing the emulation against the original is obligatory for assessing the quality of the emulation in an art context. Comparing is easier when the original hardware is still working (6.3, 6.5). An original hardware environment is also a valuable tool when tracing problems in the emulated run time environment (4.2, 5.5, 6.4).

Hence, museums should start their preservation effort while the hardware for their digital artifact is still functional and spare parts are still available.

Our digital object under preservation, the Flusser Hypertext, is an important cultural artifact that was already close to getting lost forever: only the binary files of this work still existed, stored on the physical hard disk of a vintage computer. We now have a salvaged version on hand, in the form of a disk image file that can be easily copied to various media, distributed, and placed into long-term digital storage. This version of the Flusser Hypertext is ready to execute using an "off-the-shelf" emulator.

8. ACKNOWLEDGMENTS

Thanks go to Dirk von Suchodoletz and Klaus Rechert from the University of Freiburg, who helped with the hard disk image and the setup of the emulation, and who provided their emulation framework as a bootable from stick-version. We'd also like to thank Baruch Gottlieb for establishing this contact.

The German Research Foundation DFG supported the first author with a research grant.

Adam Rosen from the Vintage Mac Museum in Boston kindly made us aware of how to overcome access restrictions to the source code of locked HyperCard stacks.

We'd also like to thank Bernd Wingert for donating his Macintosh Performa computer with the Flusser Hypertext to the Vilém Flusser Archive in Berlin, and for freely sharing his inside knowledge of the Flusser Hypertext.

9. REFERENCES

[1] Rothenberg, J. 1995. Ensuring the Longevity of Digital Information. *Scientific American*, 272(1) (Jan. 1995), 42-47

[2] Holdsworth, D., and Wheatley, P. 2001. Emulation, Preservation and Abstraction. In *Research Libraries Group RLG DigiNews* 5, 4 (Aug. 15, 2001) Online at http://sw.ccs.bcs.org/CAMiLEON/dh/ep5.html

[3] Farquhar, A., and Hockx-Yu, H. 2007. Planets: Integrated Services for Digital Preservation. *Int. Journal of Digital Curation* IJDC 2, 2 (2007), 88-99.

[4] KEEP project. Project information online at http://cordis.europa.eu/project/rcn/89496_en.pdf

[5] Cochrane, E. 2014. Emulation as a Service (EaaS) at Yale University Library. Online at http://blogs.loc.gov/digitalpreservation /2014/08/emulation-as-a-service-eaasat-yale-university-library/

[6] Brown, G. 2012. Developing Virtual CD-ROM Collections: The VoyagerCompany Publications. *Int. Journal of Digital Curation* 7, 2 (2012), 3-20.

[7] Pinchbeck, D., Anderson, D., Delve, J., Otemu, G., Ciuffreda, A., and Lange, A. 2009. Emulation as a strategy for the preservation of games: the KEEP project. In *Proc. of the Int. Conf. on Breaking New Ground: Innovation in Games, Play, Practice and Theory* (Brunel University, London, UK, Sep. 2009), DiGRA 2009.

[8] Loebel, J.-M. 2014. *Lost in Translation* [in German]. Ph.D. Dissertation. Humboldt University Berlin & VWH Verlag, Glückstadt.

[9] Matthews, B., Shaon, A., Bicarreguil, J., and Jones, C. 2010. A Framework for Software Preservation. *Int. Journal of Digital Curation* 5, 1 (June 2010), 91-105.

[10] Braud, M., Lohman, B., and van der Hoeven, J. 2012. How to run emulators remotely via the Emulation Framework. Online at http://emuframework.sourceforge.net/docs/EF-howto-remoteemulation-1.0.pdf

[11] Satyanarayanan, M. 2013. Olive: One-click Execution of Internet-Archived Software. In *NYU Scientific Reproducibility Workshop* (New York, USA, May 30, 2013).

[12] Liebetraut, T., Rechert, K., Valizada, I., Meier, K., and von Suchodoletz, D. 2014. Emulation-as-a-Service: The Past in the Cloud. In *7th Int. Conf. on Cloud Computing* CLOUD 2014. 906-913.

[13] Rechert, K., Liebetraut, T., Stobbe, O., Valizada, I., and Steinke, T. 2015. Characterization of CDROMs for Emulation-based Access. In *12th Int. Conf. on Digital Preservation* (Chapel Hill, USA, Nov. 2-6, 2015). IPRES 2015.

[14] Rinehart, R., and Ippolito, J. 2015. *Re-collection: Art, New Media, and Social Memory*. MIT Press.

[15] Kaufmann, F. 2013. Hacking Mondrian. In *Digital Art Conservation: Theory and Practice in the Conservation of Digital Art,* B. Serexhe (Ed.). Ambra V, Vienna, 273-284.

[16] Padberg, F. 2014. Emulation of Media Art or Art-Handling in the Change of Technology [in German]. In *Int. Symp. on Art-Handling* (Migros Museum for Contemporary Art, Zurich, Switzerland, Nov. 27-28, 2014).

[17] Falcao, P., Ashe, A., and Jones, B. 2014. Virtualisation as a Tool for the Conservation of Software-Based Artworks. In *11th Int. Conf. on Digital Preservation* (Melbourne, Australia, Oct. 6-10, 2014). IPRES 2014.

[18] Fino-Radin, B. 2011. Digital Preservation Practices and the Rhizome Artbase. Technical Report. Online at http://media.rhizome.org/artbase/documents/Digital-Preservation-Practices-and-the-Rhizome-ArtBase.pdf

[19] Espenschied, D., Rechert, K., von Suchodoletz, D., Valizada, I., and Russler, N. 2013. Large-Scale Curation and Presentation of CD-ROM Art. In *10th Int. Conf. on Digital Preservation* (Lisbon, Portugal, Sep. 2-6, 2013). IPRES 2013.

[20] Lurk, T., Espenschied, D., and Enge, J. 2012. Emulation in the context of digital art and cultural heritage. *Praxis der Informationsverarbeitung und Kommunikation* PIK 35, 4 (2012), 245-254.

[21] Lorrain, E. 2013. PACKED Case Study report: Mondophrenetic (2000, Herman Asselberghs, Els Opsomer, Rony Vissers). Online at http://www.scart.be/?q=en/content/case-study-report-mondophrenetic%E2%84%A2-2000-herman-asselberghs-els-opsomer-ronyvissers-0

[22] Besser, H. 2001. Longevity of Electronic Art. In *Int. Cultural Heritage Informatics Meeting* (Milano, Italy, Sep. 3-7, 2001). ICHIM 01.

[23] Irrgang, D., and Marburger, M. R. 2015. Vilém Flusser – A Biography. In *Flusseriana. An Intellectual Toolbox*, S. Zielinski, P. Weibel, and D. Irrgang (Eds.). Univocal Publishing, Minneapolis, 452-519.

[24] Böhle, K., Riehm, U. and Wingert, B. 1997. *Vom allmählichen Verfertigen elektronischer Bücher. Ein Erfahrungsbericht* [in German]. Campus Verlag, Frankfurt/M. and New York.

[25] Gottlieb, B. 2015. Hypertext. In *Flusseriana. An Intellectual Toolbox*, S. Zielinski, P. Weibel, and D. Irrgang (Eds.). Univocal Publishing, Minneapolis, 212-214.

[26] Flusser, V. 1985. *Into the Universe of Technical Images*, N. A. Roth (Transl. 2011). University of Minnesota Press.

[27] Wingert, B. 1992. Schreiben für Publizieren. Ein Hypertext-Experiment mit einem Flusser-Text [in German]. In *Kunstforum International* 117 (1992), 109-110.

[28] Wingert, B. 1991. Erfahrungen bei der Entwicklung eines Hypertextes [in German]. (talk and prototype demo). In *Arbeitskreis Literatur im Informationszeitalter* [Workshop on Literature in the Information Age] (Wissenschaftszentrum Nordrhein-Westfalen, Düsseldorf, Germany, June 20, 1991).

[29] Wingert, B. 1996. Kann man Hypertexte lesen? [in German]. In *Literatur im Informationszeitalter* [Literature in the Information Age], D. Matejovski and F. Kittler (Eds.). Campus Verlag, Frankfurt/M. and New York, 184-218.

[30] Wingert, B., 1991. (prototype demo). In *CULTEC – Kultur und Technik im 21. Jahrhundert* [Culture and Technology in the 21st Century] (Wissenschaftszentrum Nordrhein-Westfalen, Essen, Germany, Nov. 22-23, 1991).

[31] Böhle, K., Riehm, U, and Wingert, B. 1992. (prototype demo). In *Workshop Hypersystem-Konzepte in Medien und kultureller Produktion II* [Workshop on Hyper-system Concepts in the Media and Cultural Production II] (University of Lüneburg, Germany, July 13-15, 1992). Workshop program and list of exhibits online at http://www2.leuphana.de/hyperkult/archiv/hk2.pdf

[32] Wingert, B. 1992. Flusser-Hypertext. Prototyp und Entwicklungserfahrungen [in German]. (talk and prototype demo). In *GI-Symposium Hypertext und Multimedia. Neue Wege der computerunterstützten Aus- und Weiterbildung* [Symposium of the GI on Hypertext and Multimedia. Novel Approaches in Computer-supported Education], U. Glowalla and E. Schoop (Eds.). Springer, Berlin, 137-144 and 356. (Schloss Rauischholzhausen/Marburg, Apr. 28-30, 1992).

[33] Wingert, B., and Riehm, U. 1992. Wie wirken Hypertexte? [in German]. (talk and prototype demo). In *Ergebnisse der 12. Arbeitstagung Mensch-Maschine-Kommunikation* [Results of the 12th Workshop on Man-Machine Communication], S. Dutke (Ed.). Free University of Berlin, Institute of Psychology, 41-50. (Berlin, Nov. 15-18, 1992).

[34] Wingert, B. 1993. Die neue Lust am Lesen? Überlegungen zur Lesbarkeit von Hypertexten [in German]. (talk and prototype demo). In *2nd Int. Vilém-Flusser-Symposium* (Antwerpen, Belgium, Oct. 28-31, 1993). List of talks and exhibitors online at http://www.flusser-archive.org (▷Vilém Flusser ▷Symposien und Konferenzen zu Flusser ▷1993: Zweites Internationales Vilém-Flusser-Symposium)

[35] Wingert, B. 1995. Die neue Lust am Lesen? Erfahrungen und Überlegungen zur Lesbarkeit von Hypertexten [in German]. In *Kursbuch Neue Medien. Trends in Wirtschaft und Politik, Wissenschaft und Kultur* [New Media Guide. Trends in Economy and Politics, Science and Culture], S. Bollmann (Ed.). Bollmann Verlag, Mannheim, 112-129.

[36] Wiechens, P. 1998. Hypertext und Künstlerbuch. Das Buch nach dem Ende des Buches [in German]. In *Einführung in die Kulturwissenschaft* [Introduction to Cultural Studies], T. Düllo, J. Greis, C. Berthold, and P. Wiechens (Eds.). LIT Verlag, Münster, 328-346.

[37] Flusser, V. 1991. Hypertext. Über das Schicksal von Büchern [in German]. In *NZZ Folio* 10 (Oct. 1999), 35-36. Online at http://folio.nzz.ch/1991/oktober/hypertext

[38] Tögel, P. 2016. *Denk-Maschinen – Flussers Digitale Publikationen in der Ausstellung "Bodenlos – Vilém Flusser und die Künste"* [in German]. Master Thesis. Berlin University of the Arts.

Precise Data Identification Services for Long Tail Research Data

Stefan Pröll
SBA Research
Vienna, Austria
sproell@sba-research.org

Kristof Meixner
Vienna University of Technology
Vienna, Austria
kristof.meixner@tuwien.ac.at

Andreas Rauber
Vienna University of Technology
Vienna, Austria
rauber@ifs.tuwien.ac.at

ABSTRACT

While sophisticated research infrastructures assist scientists in managing massive volumes of data, the so-called long tail of research data frequently suffers from a lack of such services. This is mostly due to the complexity caused by the variety of data to be managed and a lack of easily standardiseable procedures in highly diverse research settings. Yet, as even domains in this long tail of research data are increasingly data-driven, scientists need efficient means to precisely communicate, which version and subset of data was used in a particular study to enable reproducibility and comparability of result and foster data re-use.

This paper presents three implementations of systems supporting such data identification services for comma separated value (CSV) files, a dominant format for data exchange in these settings. The implementations are based on the recommendations of the Working Group on Dynamic Data Citation of the Research Data Alliance (RDA). They provide implicit change tracking of all data modifications, while precise subsets are identified via the respective subsetting process. These enhances reproducibility of experiments and allows efficient sharing of specific subsets of data even in highly dynamic data settings.

Keywords

Data Identification, Data Citation, Reproducibility, Long Tail Research Data

1. INTRODUCTION

Human beings in general and researchers in particular are said to be lazy when tedious tasks not directly related to the primary research endeavour are due. Unfortunately, this includes providing metadata for data sets, storing, archiving and citing the data used in a research paper. Without being able to share this data, we are creating data silos, which hinder repeatability and verifyability of experiments and the re-use of data. In this work, we present three methods for improving the identification of subsets of dynamic data, by automating obnoxious tasks. Our goal is to address data identification in small and big data scenarios. Specifically, we focus on comma separated value (CSV) files, which are prevalent in both settings. Recent open data initiatives such as in the UK[1], USA[2] or Austria[3] provide access to government data for the public. Most of these portals offer a range of formats for their data sets and the majority of the formats are in plain text, allowing simple processing and human readability. More than 50 % of the data sets from the open data portals of the UK and Austria for instance are available in CSV[4]. The CSV format is used for exchanging data and often provided as data export from more complex database systems. Despite the relatively small size of individual CSV files, handling massive numbers of CSV files in multiple versions is a challenge in big data scenarios [1].

1.1 Little Science, Big Science

Contrary to many high-volume big data settings, where standardised infrastructure is available, there exist other big data settings with less mature processes, due to the lack of tools, resources and community exchange. This area is denoted the long tail of research data and subsumes large portions of data that are highly heterogeneous, managed predominantly locally within each researcher's environment, and frequently not properly transferred to and managed within well-curated repositories. The reason is that in the so called little science [2, 3], common standards and defined best practices are rare. This is particular true in research disciplines, which do not yet have a tradition of working with advanced research infrastructures and many data sets still reside on the machine of the researcher. Being able to identify which data set served as the input for a particular experiment, is based on the rigour of the scientists, and their ability to identify the particular data set again, often without proper tool support.

Reproducibility is a core requirement in many research settings. It can be tackled from several perspectives, including organisational, social and technical views. For researchers, the authors of [4] introduced 10 rules for making computational results reproducible, by describing all intermediate steps and storing the data artifacts used as inputs and produced as outputs. Both worlds - small and large scale data experiments - share the difficulty of precisely identifying data sets used as input and produced as output. This can be attributed to two main reasons: the dynamics frequently encountered in evolving data sets and tendency of scientists for using specific subsets for specific analysis that need to be precisely identified. Whereas the identification of a data set in smaller scale settings can be figuratively compared to the search of a needle in the hay stack, identifying evolving data sets in large scale environments is rather the search for the needle in a silage.

[1] http://data.gov.uk

[2] http://www.data.gov

[3] https://www.data.gv.at

[4] Data collected from the portals at 22.04.2016

1.2 Versioning Research Data

As new data is being added and corrections are made in existing data sets, we face the questions of how intermediate revisions of a data source can be efficiently managed. Having this data available, i.e. being able to obtain an earlier version of a data set, is a fundamental requirement for the reproducibility of research processes. Access to earlier versions is also essential to support comparability of experiments by running different experiments on identical data. Thus, maintaining and accessing dynamically changing research data is a common challenge in many disciplines. Storing duplicates of data sets, the prevalent approach to address this problem, hardly scales with large and distributed data volumes, while increasing the complexity of having to manage enormous amounts of data files. This calls for more efficient ways of handling versions of evolving data files.

1.3 Creating Subsets

Creating a subset is based on implicit information which records to include into a data set and which ones to omit. The basic methods needed for creating subsets are filtering and sorting. So far this process is hardly captured and researchers tend to store subsets as individual files, causing high redundancy problems and leading to an explosion of individual data files to be managed. Alternatively, researchers may chose to refer to the entire data set and provide a natural language description of which subset they were using. Albeit, this description is frequently ambiguous and may require the reader to invest significant effort to recreate the extract same subset, while making it very hard to verify whether the resulting subset is identical.

In this work we present an approach allowing to efficiently identify data sets and derived subsets, even if the data source is still evolving. These identification services can be integrated into scientific workflows, and therefore allow to unambiguously pinpoint the specific subset and version. Our approach is based upon versioning and timestamping the data as well as query based subsets of data being used in an experiment or visualisation. In our approach, we interpret *query* in a rather broad way, as by query, we understand any descriptive request for a subset of data. A query can either be an actual database query, or any operation allowing to retrieve a subset from a data source using, for example, scripts. Instead of creating duplicates of data, we use queries for (re-) constructing subsets on demand. We trace the subsetting process and assign persistent identifiers (PID) to these queries instead of static data sets. With this mechanism, we provide reproducible data sets by linking the PID to the subset creation process and matching the data against a versioned state of the source data set. This approach has been released as a recommendation by the RDA Working Group on data citation [5] and refined in [6] to address efficient and precise identification and citation of subsets of potentially highly dynamic data.

We present three implementations of this approach supporting CSV data and compare their respective advantages and disadvantages. The first approach is based on a simple file-system based versioning with script-able queries. The second approach is an extension of the first approach and based on Git branching, which enables users to work simultaneously with data sets without distracting each other. The third approach uses transparent migration of the CSV data into a relational database system, allowing more efficient versioning and more flexible query-based subset generation.

The remainder of this paper is organised as follows. Section 2 provides an overview of the state of the art from the areas of research data management, data citation and persistent identification. Section 3 outlines the challenges of dynamic data citation in research areas working with the long tale of research data. In Section 4 we introduce three realisations of the dynamic data citation method optimised particularly for small and medium-sized data sets distributed as CSV files. Section 5 provides the evaluation of the approaches, Section 6 provides the conclusions.

2. RELATED WORK

Citing publications has a century old tradition and its methods have been applied to modern scholarly communication including data sets [7, 8]. We need to be able to identify such data sets precisely. As URLs are not a long term option, the concept of persistent identifiers was introduced. Persistence is achieved by using centrally managed PID systems [9], which utilise redirection to resolve new locations of data files correctly. In many cases landing pages are the target of resolvers [10]. Landing pages contain metadata for human readers, but no standard solution regarding versioning and subsetting of data sets is provided, that is accessible by humans and machines. Recent developments try to enrich the mere redirection purpose identifier infrastructures by adding machine readable metadata [11] and providing the context of data sets [12]. We thus need to ensure that our solution can support these mechanisms of persistent data identification and citation by allowing the assignment of PIDs to data.

Current citation practices usually refer to static data files. However, we increasingly find situations where such data files are dynamic, i.e. new data may be added at certain intervals of time. For working with the data as it existed at a specific point in time (e.g. to verify the repeatability of an experiment, or to compare the result of a new method with earlier published results), we need to ensure that we can provide exactly the same data as input. To achieve this, data needs to be versioned, which is a common task in the data management domain [13] and implemented in software applications dealing with critical data [14]. With decreasing storage costs preserving previous versions even of high volume data has become a service offered by many data providers but still storing multiple versions is a challenge [15]. Storing previous versions of data sets is usually accompanied by timestamps [16]. Each operation which changes the data is recorded and annotated with a timestamp of its occurrence.

As mentioned above, natural language description frequently is not precise enough to unambiguously describe a specific subset. Storing redundant copies, on the other hand, does not scale well. Thus, the concept of a dynamic identification of subsets using query stores has been introduced [17]. The query store does not only store the queries as text, but also preserves the parameters of each query. This allows providing this information on other representations than the original query and enables to migrate the query to other systems. The query store operates on versioned data and queries [18], which allows retrieving only those versions of the records which have been valid during the original execution time. The data and the queries are versioned, the system can be used for retrieving subsets of large data sources exactly the same way as they have been at any given point in time [19].

The Research Data Alliance (RDA) Data Citation Working Group published 14 recommendations on how to make research data citable [6]. The RDA data citation mechanism can be used for evolving and for static data and is based upon versioned data and query mechanisms, which allow to retrieve previous versions of specific data sets again.

3. CHALLENGES IN HANDLING SMALLER SCALE RESEARCH DATA

Research disciplines and data in the so-called “long tail” are often suffering from a lack of professional tools and infrastructure which could support researchers in creating, using, sharing and verifying research data sets. If peer researchers want to repeat the process again or reuse the data to compare results of different approaches on the same data, means for verifying if the correct data were used are essential. Yet, this is far from trivial, with complexity caused primarily by two issues: dynamics in the data and the usage (and thus precise identification) of specific subsets of data.

3.1 Versioning Approaches: How Change is Traced

While researchers used to share static data files in the past, in current research settings the data we use is increasingly dynamic: new data being added, errors being corrected, wrong items being deleted. Ways this is dealt with include batch release of subsequent versions of the data, resulting in delayed release of corrections as they need to be aggregated until the next monthly, quarterly or annual release is due, as well as managing many redundant files, leading to high complexity in file naming and versioning conventions. Typically researchers utilise a rename and copy approach, where each version of a data set is distinguished by its file name. Recommendations for naming files exist [20], suggesting to use project or experiment name or acronym, coordinates, names, dates, version numbers and file extension for application-specific files. Nevertheless it is cumbersome and error prone for researchers. We thus need automated procedures allowing researchers to manage different versions of evolving data, allowing them to go back to earlier versions of data when needed. This should happen in an automated way, not putting the burden of version management and identification of changes on the researcher.

3.2 Creating Subsets From Implicit Information

Researchers often work with subsets from larger data sources, for curating specific aspects of a data set or visualising a specific view. Many publications only cite the full, raw data source and describe used subsets only superficially or ambiguously, by using natural language description for instance. From a reproducibility perspective, it is essential to know precisely, which subsets of data was used during a processing step. In contrast to large scale systems, which often guide researchers through standardised workflows of data filtering, the procedures in smaller scale research are often less well structured and defined. For this reason there is a larger variance in the way how subsets of data can be obtained and how subsets have been created. In larger scale settings, sophisticated database management systems are in place. In the small scale domain, text processing or spreadsheet programs are often used for creating a subset from a file. Scripting languages allow filtering, sorting and selecting subsets from file in a more automated way, but obtaining a specific subset again from a versioned data file in a reproducible way is a challenge.

For making implicit sub-setting information explicit, we need to trace the subset creation process itself and store this information in a persistent way. As manual work is susceptible to errors, an automated solution is a basic requirement for the integration of identification as a service into existing scientific workflows.

4. PRESERVING THE INFORMATION OF THE SUBSET CREATION PROCESSES

For this reason we introduce three implementations for the automated, unique identification of data, based on the data citation concepts introduced by [17, 19] and on the RDA recommendations for data citation [6]. This dynamic data citation is based upon two generic principles: (1) In order to be able to retrieve earlier versions of data, the underlying data source must be timestamped and versioned, i.e. any addition to the data is marked with a timestamp, and delete or update of a value is marked as delete and re-insert with respective timestamps. (2) Subsets are identified via a timestamped query, i.e. the query that was executed to generate a subset is retained in a so-called query store to enable its re-execution with the according timestamp. By assigning persistent identifiers to this information, understanding and retrieving the subset at a later point in time is possible. Integrating a query based identification service improves the reproducibility of scientific workflows in small and large scale scientific research likewise.

The three implementations of these principles for CSV files presented below differ primarily in their way of storing the data in the back end. Two approaches are based on Git, a wide spread version control system for source code, one approach utilises a migration process into a database system. The first approach uses a simple versioning scheme (Git) leading to low system complexity, but also less flexibility in subset generation and lower scalability. The second approach is also based on Git and utilises the branching model allowing simultaneous editing of data sets. The third approach migrates the CSV file transparently into a relational database, leading to higher complexity in system maintenance but providing higher efficiency and flexibility. In all three cases, the subset is identified via the query mechanisms (i.e. database queries via an API or graphical interface, scripting languages or via scrip-table SQL statements operating on the CSV file). The queries used in all three approaches are timestamped and associated with a PID. It is worth noting that we utilise a simplified PID approach in this paper, but the principle is compatible with accepted solutions such as DOI or other PID systems.

4.1 Using Git for Creating Reproducible Subsets

Source code management software and distributed revision control systems such as Git[5] or Subversion[6] are spreading from the software development departments to the labs,

[5] https://git-scm.com/

[6] http://subversion.apache.org/

as version control systems allow working collaboratively on files and trace changes. These systems have been designed for plain text file formats, as their change detection algorithms are based on the comparison of strings. If each change of a file is committed into the repository, the changes are traceable and previous versions of each can be compared with the current revision.

Many different tools exist for manipulating CSV data, ranging from command line applications such as awk, sed, csvkit[7] to scriptable statistical software such as R.

In the following example use case, users provide a list of the Top500[8] super computers in a CSV file as input for the script. The list is updated periodically and each change is committed to the Git repository. A user interested in analysing the top-50 computers for some study creates an according subset and selects the columns Rank, Site and Cores from the file. The subset will be stored in the location provided as the second parameter to the script. Listing 1 provides a simple example for creating such a subset of CSV data using the mathematical software R. Listing 2 shows the execution of the script in a Linux shell.

Listing 1: Rscript for Subsetting

```
# Create a subset of the
# top 5 of the Top500 list
args <- commandArgs(trailingOnly = TRUE)
inputDatasetPath=args[1]
outputSubset=args[2]
dataset <- read.csv(inputDatasetPath,
    header=TRUE)
subset <- subset(dataset,
    Rank<=5,select=c(Rank,Site,Cores))
write.csv(subset, file=outputSubset)
```

Listing 2: Executing the Script

```
# Execute the R script and
# obtain a subset from the provided CSV file
/usr/bin/Rscript top5-subset.r \
/media/Data/Git-repository/ \
    supercomputing/supercomputer.csv \
/media/Data/Git-repository/ \
    supercomputing/supercomputer-top5.csv
```

We store these scripts in Git to retrieve the very same data set again, by executing the proper version of a script against the correct version of the data set. To do so, we store the CSV file name and location and the execution timestamp in a metadata/landing page file in the Git based query store. Each query is assigned a PID, which serves as file name of the according metadata file in the query store, which allows retrieving the data later by resolving the PID to the file name.

We implemented a prototype based on the Eclipse JGit Java library[9], which offers a low level API for the interaction with Git repositories. Revisions of the data set are committed to the repository, where Git stores a commit hash and the timestamp of the update. If users want to retrieve a subset again at a later point in time, they first retrieve the metadata file from the Git system using the PID as the file name. This file then provides the file name of the CSV data set and the execution timestamp of the query. In the next step, the system traverses the revision tree with the RevWalk object and builds a revision graph based on the commit dates[10]. We filter the commits and select the closest timestamp valid before the execution of the script[11]. This revision was valid during the execution of the original query. We fetch this version from the repository and re-execute the R script against the versioned data set, as depicted in Figure 1.

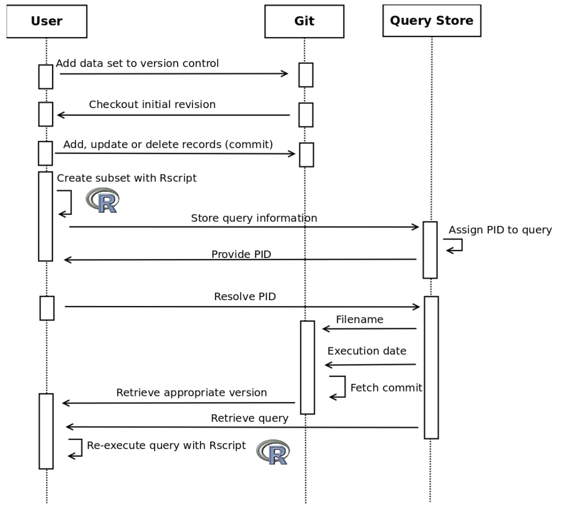

User
Git
Query Store
Add data set to version control
Checkout initial revision
Add, update or delete records (commit)
Create subset with Rscript
Store query information
Assign PID to query
Provide PID
Resolve PID
Filename
Execution date
Fetch commit
Retrieve appropriate version
Retrieve query
Re-execute query with Rscript

Figure 1: The CSV Subsetting Workflow with Git

For making this process reproducible, the user commits both, the CSV data file and the R script into the Git repository. The metadata files are committed into the Git repository in a separate PID folder. This folder contains all PID identified metadata files of reproducible data sets, using the PID as the file name. This allows us establishing a unique link between the PID and the metadata file, and by the transitivity, also with the data and the scripts. The metadata file contains the execution time, application version, the script and its parameters used as well as the re-execution steps for each subset. The metadata required can be generated automatically by using Git tools, no additional software dependencies are required. Listing 3 shows an example for the collected metadata and the references to versioned data and script files.

Listing 3: The Metadata File

```
# PID=1234/abcdefgh
# Repository_Path=/media/Data/Git-Repository
# Execution_Time=2015-09-30:11:07:09
# Subset_Tool=R scripting front-end version 3.2.2
    (2015-08-14)
# Subset_Tool_Path=/usr/bin/Rscript
# Input_Script_Path=supercomputing/top5-script.r
# Input_Script_Hash=bef5d...d7861:supercomputing/top5-
    script.r
# Dataset_Path=supercomputing/supercomputer.csv
# Dataset_Commit_Hash=acaed...4cf9c:supercomputer.csv
# Output_Path=/tmp/supercomputer-top5.csv

# Original execution:
#   /usr/bin/Rscript supercomputing/top5-script.r \
#   /media/Data/Git-repository/supercomputing/
    supercomputer.csv \
#   /tmp/supercomputer-top5.csv

# Recommended re-execution
# Retrieve script
git --git-dir=/media/Data/Git-Repository/.git/ \
show bef5d...d7861:supercomputing/top5-script.r \
> /tmp/reproduced-datasets/top5-script.r
# Retrieve data set
git --git-dir=/media/Data/Git-Repository/.git/ \
show 47bed...b9792:supercomputing/supercomputer.csv \
> /tmp/reproduced-datasets/supercomputer.csv
# Rexecute
/usr/bin/Rscript supercomputing/top5-script.r \
/tmp/reproduced-datasets/supercomputer.csv \
/tmp/reproduced-datasets/supercomputer-top5.csv
```

[7] http://csvkit.readthedocs.org/en/0.9.1/

[8] www.top500.org

[9] https://eclipse.org/jgit/

[10] Code snippet: https://gist.github.com/stefanproell/b38e496a1259472c75f0

[11] Code snippet: https://gist.github.com/stefanproell/34f8ac3fb5b63599976f

The method we propose is a simple way of storing reproducible data sets within Git repositories. The format of the metadata file serves as documentation and is machine actionable, as it allows retrieving the subset by executing the script file. The metadata can be parsed and used in a landing page, for increasing the readability for human users. It works well for simple scripts, which are not depending on processing chains with user interactions. It is designed to support one user per time per data set and implements an evolution pattern for each data set.

Note that in order for this approach to work, the repository has to ensure that the access/scripting language used to identify the subset is maintained. We thus recommend to only support subsetting functionality with a clearly and unambiguously defined semantic. All complex processing (e.g. data analysis, visualisation, etc.) should happen in subsequent processing scripts to keep the complexity of the long-term stability manageable. Considering more complex scenarios blurs the border between reproducible data sets and process preservation.

In addition to the R-based (or, in fact, any similarly structured script-like interface) we also provide support for subsetting using an SQL-like query language that can be executed against CSV files via the CSV2JDBC[12] library for Java, which allows retrieving subsets from CSV files via SQL statements. As both CSV and SQL are based on a tabular view of the data, CSV data can be easily mapped into a relational database table. Hence the translation process of a CSV subset selection process can be mapped to an SQL query. Figure 2 shows this transition.

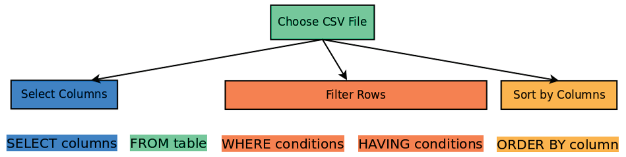


Figure 2: CSV Subsetting and SQL Queries

When a user wants to create an identifiable subset, we store the selected columns, the filter parameters and the sorting information in the query store. We preserve the SQL statement used for obtaining the subset in the first place. Additionally, we store the CSV file name and location and the execution timestamp in a metadata/landing page file, also stored in the Git based query store. As each metadata file has the unique PID as file name, the query can be re-executed based on the versioned CSV data set.

4.2 Using Git Branching to Separate Data and Queries

In Section 4.1 we introduced an approach for storing CSV data and metadata files in different folders in a Git repository. Furthermore we explained how to retrieve the metadata and CSV data files in order to re-execute the queries on the subsetted data. In this Section we will present a second approach that brings several advantages in a collaborative work environment.

When working with Git in a shared environment the concept of branching is the recommended best practise for allowing multiple researchers to work with different states of the data or files at the same time. A branch allows researchers to work with a specific version of data (or files) without distracting others. After the work has been completed (e.g. a subset has been created), the data can be merged with the main line or other branches again. At a certain point these branches are then merged together to a single branch to generate a common state.

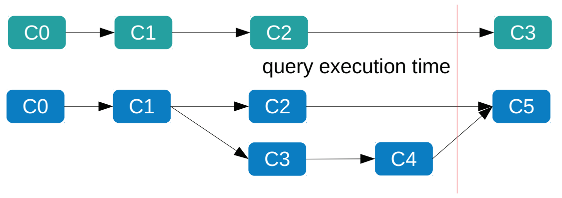


Figure 3: Commit graph without and with branches

Figure 3 shows two commit graphs. The upper graph represents commits to the repository done on a single branch, as described in Section 4.1. The graph below represents a repository were after commit `C1` a second branch was opened. The subsequent commits `C3` and `C4` were committed to the second branch. Commit `C5` is a merge commit where the two branches are merged together to a single branch.

If a query was executed at the time, that is marked by the red arrows in Figure 3, the algorithm introduced in Section 4.1 works differently on the two graphs, if it is re-executed at a time after commit `C5`. In the repository represented by the upper graph the algorithm returns the correct commit `C2`. In the repository represented by the lower graph the query would return commit `C4` instead of `C2` because it has a later commit date.

To solve this issue we need to change two aspects of the prior solution. First we need to save the CSV and metadata files in two separated branches instead of different directories. Second we change the algorithm to retrieve the data based on the timestamp to an approach were the specific commit hash is used.

If the CSV and metadata files are saved in the same branch, as in the approach described above, the history of CSV commits would be cluttered by the commits of the metadata. We therefore create a dedicated branch for the data and the queries[13].

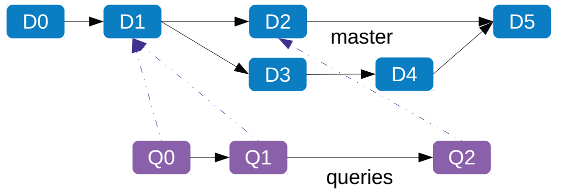


Figure 4: CSV(*master*) and metadata branch (*queries*)

Usually branches in Git share a common commit as ancestor, which means that the branches also share the same

[12] http://csvjdbc.sourceforge.net/

[13] https://gist.github.com/Mercynary/cc0d6bee550b701ab3c2ae6add75aee8

history up to this point. In Figure 3 the common ancestor is labeled as C1. Git also supports orphaned branches, where the the diverging branch gets a new commit as starting point for the history. The commit graphs of a repository with this configuration are displayed in Figure 4. In order to clearly separate the branches and their history, we create the query branch as an orphaned branch.

With the two branches created we can now change the algorithm to store the metadata files and retrieve the datasets on which the queries are executed on[14]. Figure 4 should thereby serve as an example of a repository with a *master* and a *queries* branch, as well as a branch that contains two CSV file commits on a third branch that was merged at a point later in time than the query commit Q2. In the figure the commits Q1 and Q2 refer to the CSV file commit D1 as expected. However, the query commit Q2 refers to the CSV file commit D2 as the third branch was not visible to the application at the time of the query execution. In the description below the labels of the commits also represent the hash values of the commits.

Firstly, for saving a metadata file in the repository, the PID provided by the user is hashed with *SHA-1* to a string that can be used as a file name. We do this because PIDs could contain special characters, that are not permitted in a file name. Although we are aware of the probability of hash collisions, we chose *SHA-1* because Git uses the same algorithm to calculate the hashes of the committed files and thus limits our approach in the same way.

Secondly the previously created query branch is checked out and the contents of the metadata file are written. The most important information saved in the metadata is the query and its parameters, the data file name, the PID and the commit hash of the latest revision of the data branch. The commit hash is sufficient to identify the commit and its commit time in Git as well as to locate it on a specific branch if necessary, as the hashes remain the same when the branches are merged at a later point in time. In case of Figure 4 the saved commit hash in the metadata file would be D2. In a last step the metadata file itself is committed to the query branch which results in commit Q2. In this way the structure and history of the CSV data branch and the metadata branch do not interfere with each other.

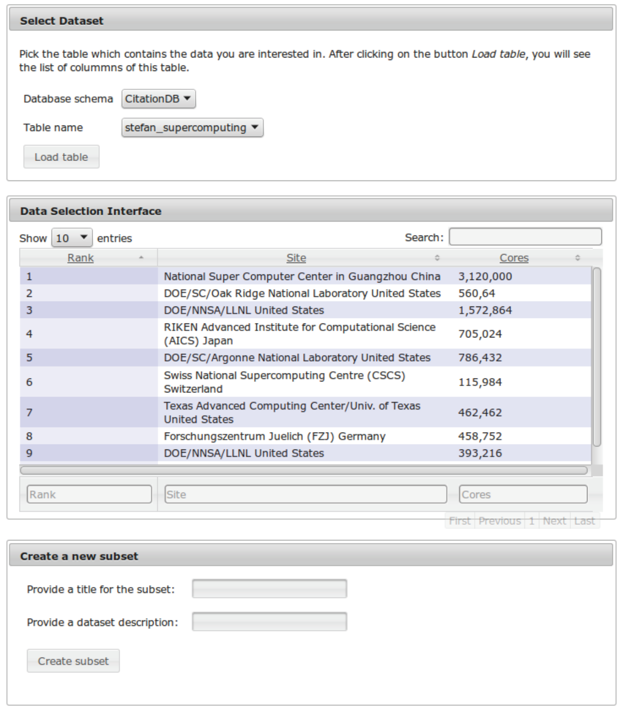
To retrieve the queries and re-execute them on the correct dataset, the following algorithm needs to be applied[15]. The user first provides the PID via the web application which is then hashed to get the file name of the metadata file. The next step is to checkout the query branch and read the metadata file identified by the hashed PID. From this metadata the commit hash and file name of the CSV data file can be extracted. In the example depicted in Figure 4 we would get the hash value D2. We then checkout the exact commit that is identified by the hash value. This way we restore the CSV data file as it was at the time the query was executed the first time. At this point we then can re-execute the query on the dataset.

Because the metadata files store the unique commit hash of the CSV data file in the repository, at the time when the query is stored and executed, the commits can not get mixed up when two or more dataset branches are merged together in advance to the approach that was based on timestamps. As mentioned in the beginning of this section, due to the solution of separated CSV data and matadata file branches as well as a metadata retrieval based on commit hashes this approach is better suited when working in a collaborative environment. We implemented a prototypical web application[16] as a proof of concept.

4.3 Reproducible Subsets Based on Migration and Database Queries

One major disadvantage of CSV files is the lack of native support for subsetting, i.e. selecting only a specific set of rows and columns. While the Git approach is suitable for smaller scale files, native support is preferable for larger data files, allowing to extract only smaller files from a repository in the first place, rather than having to extract all data and performing the subsetting afterwards.

Our third implementation still transparently provides CSV files for the researcher, but internally utilises the advantages of a database management system which takes care of versioning the data. Users upload their CSV files into a Java application and they can generate a subset by using a Web interface. Subsets can be retrieved again as CSV file download, where the files are created on demand based on queries. We implemented a two phased migration process for inserting the data into a MySQL 5.7 database management system. Figure 5 shows the interface of our prototype solution with three selected columns.



Select Dataset

Pick the table which contains the data you are interested in. After clicking on the button *Load table*, you will see the list of colummns of this table.

Database schema: CitationDB

Table name: stefan_supercomputing

Data Selection Interface

Show 10 entries

Search:

Rank	Site	Cores
1	National Super Computer Center in Guangzhou China	3,120,000
2	DOE/SC/Oak Ridge National Laboratory United States	560,64
3	DOE/NNSA/LLNL United States	1,572,864
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	705,024
5	DOE/SC/Argonne National Laboratory United States	786,432
6	Swiss National Supercomputing Centre (CSCS) Switzerland	115,984
7	Texas Advanced Computing Center/Univ. of Texas United States	462,462
8	Forschungszentrum Juelich (FZJ) Germany	458,752
9	DOE/NNSA/LLNL United States	393,216

Create a new subset

Provide a title for the subset:

Provide a dataset description:

Figure 5: An Interface for Creating Reproducible Subsets

In the first phase, the CSV file parsed and a table schema based on the file structure is created. CSV header names (i.e. the first row in the CSV file) serve as column names for the table. In cases where a CSV schema[17] file is available, the data type can be specified for the columns within the database table. If no schema is available, the data in each column can be analysed and heuristics can be used to determine an appropriate data type (date, numeric, string, etc.), or all columns can simply be interpreted as text strings (VARCHAR columns). By parsing the file once, columns containing potential identifiers can be detected. We use these identifiers as primary keys for differentiating between records. If no candidate is available, this can either be an indicator for duplicate records in a CSV file or the set simply does not provide unique columns which could serve as identifiers, a sequence number column generated automatically by the system is appended to the data set for internal use. Each newly generated table is expanded by one column for timestamps and one column for storing the event type (*INSERTED*, *UPDATED*, *DELETED*). Having the two additional columns available allows implementing a versioning scheme as described in [18]. In the second phase, the CSV file is read row by row and the data is inserted into the database table. For each newly added record, the system automatically inserts the timestamp and marks the data as inserted.

For adding data to the set, users can provide CSV files with the same structure and upload them into the system. Header names can serve for checking whether the file still has the same structure and the column type heuristics can be applied for checking if the column type remained the same. During the upload, the file is parsed and the records are inserted into the data set, where the primary key column defined in the database ensures that updates can be detected.

Upon the upload of a file containing changes, old records are marked as updated and the updated version of that record is inserted as a new version with the current timestamp and the *INSERTED* mark. Obviously, detecting which record to update only works if a primary key is present in the updating file. In case where no such unique column is available, researchers can download the current version of the data set including the sequence number column. By updating this file, for instance by using some spreadsheet software, the sequence number can be preserved and used as a matching column.

The query store is implemented as a separate database schema, providing tables for storing the metadata for retrieving the queries at a later point in time. The query metadata includes source table, query parameters, execution time and the persistent identifiers assigned to the query. As soon as the data has been migrated into the RDBMS, the advantages of the query based mechanism can be used for identifying specific subsets of research data. This allows to re-execute the query and map the timestamp of the query execution time against the versioned data set. The subset which is defined by the information stored in the query store can then be retrieved on demand.

5. EVALUATION OF THE DATA CITATION APPROACHES FOR LONG TAIL RESEARCH DATA

In this paper we presented three approaches for creating reproducible and citable subsets of CSV data. All three are based on versioned and timestamped data and utilise a query mechanism which retrieves the data at it was at a specific point in time.

5.1 Using Git for Versioning, Identification and Re-Execution

Versioning data sets with Git is easy to integrate and commonly recognised as good practice for text based data formats. The overhead created by the Git repository is low and does not require sophisticated server infrastructure. Interpreting software scripts as query allows to create subsets in a flexible way. Instead of adding subsets directly into the Git repository as new files, the query string or script can be used for retrieving the data from the versioned data set. The query or scripts respectively are versioned as well and thus can be mapped to a specific version of a subset. As the version of the data set can be obtained from the repository, the likewise versioned query can be re-executed without any modifications. The mechanism can be applied to any scripting language, as long as the required commands and parameters are stored in the query store.

By adding the script files responsible for creating the subsets under version control, internal mechanisms of Git can be used to re-execute the subset creation process at any timestamp. Various versions of the script can be applied on the available history of data sets. This does not only enable reproducibility, but also allows to compare different versions of the subsetting process with each other.

The required software is open source and thus freely available and used by a very large community. The Git based approaches can therefore easily be implemented in long tail data settings. Furthermore it can be integrated into existing processing pipelines, adding reproducibility for the data input and output processing steps.

Git utilises a line based approach for interpreting differences in versions of data. Thus the traceability of changes between two versions is limited, if the granularity is below row level. Sorting for instance can hardly be differentiated from updating records, which results in the deletion and subsequent addition of a record into the file.

Re-ordering a CSV file by changing the sequence of columns also leads to a completely different file, as all of the records are considered as deleted and new records are detected to be added. For this reason different versions of one data set cannot be compared reliably without additional tools, leading to less-efficient utilisation of storage. On the other hand, as CSV files tend to be moderately-sized, this does not constitute a major limitation.

Similarly, for retrieving a subset, the entire CSV file first has to be checked out of the repository before the appropriate subset can be extracted by running the original script. While this might be undesirable in massive-scale data settings it is unlikely to cause major problems in typical settings employing CSV files.

These limitations of the Git based approaches are due to the focus of Git on source code rather than data files. The Git approaches allow utilising one single versioning system for both, code and data. Therefore, no complex infrastructure or maintenance is required and the integration of the data citation solution into existing workflows suitable for any kind of ASCII data files and scripting languages for

[14] https://gist.github.com/Mercynary/bac394c035b0a98b338202d3e8705768

[15] https://gist.github.com/Mercynary/8703ea098bcec52a2233537aaecc9f20

[16] https://github.com/Mercynary/recitable

[17] http://digital-preservation.github.io/csv-schema/csv-schema-1.0.html

retrieving the subsets requires low overhead. Subsets can be compared across different versions by creating delta files (also known as diffs) and the differences can be visualised or extracted.

5.2 Using Database Systems for Versioning, Identification and Re-Execution

Advanced database technologies support very large data sets and provide a higher performance than the file based approaches. Flexible query languages such as SQL allow retrieving specific subsets from data. Using a graphical interface hides the complexity of the query language and users can select and re-order columns in the data set, filter and sort the rows according to specific criteria, much as they are used to work with data in spreadsheet programs.

Rewriting the queries for retrieving the version valid at a specific timestamp is a necessity, but can be automated by intercepting the commands from the interface. By using triggers all of the required operations can be automated and indices on the metadata columns increases the query performance for versioned data. In our approach, we store the filter and sorting criteria from these queries in the query store. Thus we can re-execute any query at any available point in time on the versioned data and provide additional services on top such as result set verification by hashing.

The database approach overcomes the limitations of the Git based data citation approach as it does not rely on line level versioning, but allows fine granular data citation even for single cells in a table. As SQL is a declarative language, the subsets of data are described in an abstract way, which allows domain experts to understand what a specific query returned as result. This information can be represented in different forms and can be reused for instance for providing automated citation text snippets automatically.

Subsets can be compared across different versions, simply by re-executing the stored query with different timestamps. Differences can be made visible by comparing the returned result sets and exporting the differences. Handling alternative sortings or a different sequence of the columns of a data set can be easily handled by rewriting queries, without the need of changing the underlying data set.

The flexibility offered by this approach comes with the cost of increased complexity. The data needs to be imported into the database system, which is responsible for versioning both, the data and the queries including their metadata. Also the interface needs to support users depending on the requirements of the domain, therefore the solutions may often not be applicable as a generic tool across community boundaries.

5.3 Preserving the Subset Creation Process for the Long Term

Both approaches for data citation allow researchers to generate precisely defined subsets from CSV files. The processes we described provide reproducibility for data sets, as they tie the versioned data and a timestamped query together. In contrast to storing the different versions of a subset as individual files, the processes require less disk space and the information how a subset was created is maintained inherently, as this information is contained in the query. This allows domain experts to understand what records have been included into a subset and which ones have been omitted. As all parameters of the query are stored, either in the query store explicitly or within a script, the subset creation process can be understood ex post. This knowledge contributes not only for the long term reproducibility, but also adds valuable metadata automatically, as the intention of a subset can be traced even if the data should be long gone.

While both approaches are simple and easy to implement, they both rely on the availability of the query language (i.e. SQL, software scripts, etc.) and the execution environment of the query engine. In order to keep these subsets accessible, the environment and the query engine need to be preserved for the long term. As technology progresses, the emulation of the original environment or the migration to a new environment may become necessary. Our approach is based on two de-facto industry standards: Git and SQL. For source code versioning, migration pathways to new versions of the Git software or other source code management systems exist already today. The same applies to the database system MySQL, which offers many migration pathways to other (relational) database systems and even back to CSV. Whenever one of the two system needs to be moved to a virtualised environment or migrated into a different environment, the correctness of the re-executed subsets needs to be verified. This can be achieved by comparing the hash values from the old and the new system.

6. OUTLOOK AND CONCLUSIONS

In this paper we present three methods for the precise identification of arbitrary subsets of CSV files even when these data files are evolving over time. The three methods have in common that they allow to make specific subsets of data citable, by assigning a PID to the subsetting process ("query") executed against a specific state (timestamp) of the versioned data source. Additionally, we store the query or script respectively, which created the subset in a versioned fashion. We establish a link between the versioned data set at a specific time and the query as it was executed at that point in time. Being able to reproduce the process of subset creation allows us to shift the identification mechanism from data set level to the query level. This produces much less storage overhead as the duplication of data is avoided. Storing query metadata does not require significant storage compared to versioned subsets of data.

The solutions we propose have been developed with a focus on simplicity, low overhead, low maintenance and the ease of use in various research settings. The steps necessary to create citable subsets can be fully automated, relieving the researcher from the burden of individual data management, i.e. manually maintaining multiple copies of data files. The approaches can be used in combination with a centralised repository or individually at the researchers work station.

The first two approaches rely on an underlying Git repository to be used for data storage and for providing versioning capabilities in long tail research data settings. The subsetting is performed by scripts which create a subset from a data set. Both, the data and also the scripts required for creating the subset are stored in a Git repository. Additional metadata allows to re-create a subset as it was at any given point in time. Descriptive information can be added, which allows human beings to understand how a subset was created which further improves the reproducibility of data driven experiments. The first approach is simplistic, equipping researchers with a simple yet powerful method for creating citable data sets, by storing the data and the script in a dedicated repository in a linear fashion. Each subset becomes identifiable with a PID. The second approach adds parallelism to the approach and allows several researchers to simultaneously work with data sets without distracting each other. The results can be compared and easily shared.

In the third implementation, the CSV data is migrated into a relational database. Subsets can be generated either directly via an API accepting SQL queries, or via a graphical interface mimicking a spreadsheet program. By storing the data as well as the subsetting information in a versioned fashion in a database system, subsets from very large data sets can be made citable in a efficient way. Additionally, the proposed methods allow comparing different versions of the same subset more easily and allow generating subsets with the same characteristics also from newly added data. Storing the query allows retrieving in fact any subset version of evolving data and enhances the reproducibility of data driven research in larger scale settings.

Acknowledgement

Part of this work was co-funded by the research project DEXHELPP, supported by BMVIT, BMWFW and the state of Vienna, and COMET K1, FFG - Austrian Research Promotion Agency.

7. REFERENCES

[1] Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen, et al. The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing*, 2012.

[2] Derek John de Solla Price, Derek John de Solla Price, Derek John de Solla Price, and Derek John de Solla Price. *Little science, big science... and beyond*. Columbia University Press New York, 1986.

[3] Christine L Borgman, Jillian C Wallis, and Noel Enyedy. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries*, 7(1-2):17–30, 2007.

[4] Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. Ten simple rules for reproducible computational research. *PLoS Comput Biol*, 9(10):e1003285, 10 2013.

[5] Andreas Rauber, Ari Asmi, Dieter van Uytvanck, and Stefan Proell. Data Citation of Evolving Data - Recommendations of the Working Group on Data Citation. https://rd-alliance.org/rda-wgdc-recommendations-vers-sep-24-2015.html, September 2015. Draft - Request for Comments.

[6] Andreas Rauber, Ari Asmi, Dieter van Uytvanck, and Stefan Proell. Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use. *Bulletin of IEEE Technical Committee on Digital Libraries*, 2016. Accepted for Publication. URL: https://rd-alliance.org/system/files/documents/RDA-Guidelines_TCDL_draft.pdf.

[7] Heather A Piwowar and Todd J Vision. Data reuse and the open data citation advantage. *PeerJ*, 1:e175, 2013.

[8] Mark P Newton, Hailey Mooney, and Michael Witt. A description of data citation instructions in style guides. 2010.

[9] Jochen Kothe Hans-Werner Hilse. *Implementing Persistent Identifiers: Overview of concepts, guidelines and recommendations*. Consortium of European Research Libraries, London, 2006.

[10] Ruth E Duerr, Robert R Downs, Curt Tilmes, Bruce Barkstrom, W Christopher Lenhardt, Joseph Glassy, Luis E Bermudez, and Peter Slaughter. On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics*, 4(3):139–160, 2011.

[11] Norman Paskin. Digital Object Identifier (DOI) System. *Encyclopedia of library and information sciences*, 3:1586–1592, 2010.

[12] Tobias Weigel, Michael Lautenschlager, Frank Toussaint, and Stephan Kindermann. A framework for extended persistent identification of scientific assets. *Data Science Journal*, 12(0):10–22, 2013.

[13] John F Roddick. A survey of schema versioning issues for database systems. *Information and Software Technology*, 37(7):383–393, 1995.

[14] R. Chatterjee, G. Arun, S. Agarwal, B. Speckhard, and R. Vasudevan. Using data versioning in database application development. In *Software Engineering, 2004. ICSE 2004. Proceedings. 26th International Conference on*, pages 315–325, May 2004.

[15] Divyakant Agrawal, Amr El Abbadi, Shyam Antony, and Sudipto Das. Data management challenges in cloud computing infrastructures. In *Databases in Networked Information Systems*, pages 1–10. Springer, 2010.

[16] Peter Buneman, Sanjeev Khanna, Keishi Tajima, and Wang-Chiew Tan. Archiving scientific data. *ACM Trans. Database Syst.*, 29(1):2–42, March 2004.

[17] Stefan Pröll and Andreas Rauber. Citable by Design - A Model for Making Data in Dynamic Environments Citable. In *2nd International Conference on Data Management Technologies and Applications (DATA2013)*, Reykjavik, Iceland, July 29-31 2013.

[18] Stefan Proell and Andreas Rauber. A Scalable Framework for Dynamic Data Citation of Arbitrary Structured Data. In *3rd International Conference on Data Management Technologies and Applications (DATA2014)*, Vienna, Austria, August 29-31 2014.

[19] Stefan Pröll and Andreas Rauber. Data Citation in Dynamic, Large Databases: Model and Reference Implementation. In *IEEE International Conference on Big Data 2013 (IEEE BigData 2013)*, Santa Clara, CA, USA, October 2013.

[20] Matthias Schwab, Martin Karrenbach, and Jon Claerbout. Making scientific computations reproducible. *Computing in Science & Engineering*, 2(6):61–67, 2000.

CERN Services for Long Term Data Preservation

Frank Berghaus, Jakob Blomer,
Germán Cancio Melia,
Sünje Dallmeier Tiessen,
Gerardo Ganis,
Jamie Shiers, Tibor Simko
CERN
1211 Geneva 23, Switzerland
+41 22 76 76111
{ Frank.Berghaus, Jakob.Blomer,
German.Cancio.Melia,
Sunje.Dallmeier-Tiessen,
Gerardo.Ganis, Jamie.Shiers,
Tibor.Simko }@cern.ch

ABSTRACT

In this paper we describe the services that are offered by CERN [3] for Long Term preservation of High Energy Physics (HEP) data, with the Large Hadron Collider (LHC) as a key use case.

Data preservation is a strategic goal for European High Energy Physics (HEP) [9], as well as for the HEP community worldwide and we position our work in this global content. Specifically, we target the preservation of the scientific data, together with the software, documentation and computing environment needed to process, (re-)analyse or otherwise (re-)use the data. The target data volumes range from hundreds of petabytes (PB – 10^{15} bytes) to hundreds of exabytes (EB – 10^{18} bytes) for a target duration of several decades.

The Use Cases driving data preservation are presented together with metrics that allow us to measure how close we are to meeting our goals, including the possibility for formal certification for at least part of this work. Almost all of the services that we describe are fully generic – the exception being Analysis Preservation that has some domain-specific aspects (where the basic technology could nonetheless be adapted).

Keywords

Data reuse; preservation; reproducibility; research data management; digital curation; virtualization; digital libraries; massive storage; open access; open data.

1. INTRODUCTION

CERN, the European Centre for Nuclear Research, is situated outside Geneva with much of its facilities spreading into neighbouring France. It has existed for over 60 years and has 21 member states with several more in various preparatory stages of membership. CERN performs research into the nature and structure of the Universe – the fundamental particles, e.g. the constituents of the constituents of the constituents of atoms[1], and the forces that act between them.

CERN has a diverse research programme based on a wide range of particle accelerators. The largest and most powerful of these is the LHC that entered production in late 2009 after many years of preparation. The LHC occupies a circular tunnel some 100m underground and 27km in circumference. The tunnel previously housed the Large Electron Positron collider (LEP) that operated from 1989 – 2000 and the data from LEP are still available and actively used. Although the total data volume from LEP is "only" around 500TB, this was also "Big Data" in its day. Also, lessons learned from LEP – where we expect the data to be usable up to around 2030 – point the way to what we can expect to achieve for the LHC.

Unlike many disciplines that make *observations* – by definition unrepeatable – HEP makes *measurements*[2]. It would, for example, be technically possible to build a new LEP machine and data from such a machine could make the existing data entirely redundant. This is a simplifying characteristic that is not shared, for example, by gravitational wave detectors, space telescopes and earth observing systems: there, if an event is missed or poorly measured, it can never be measured again – the opportunity has been lost for eternity.

The four main detectors at the LHC, named after the corresponding worldwide scientific collaborations with up to several thousand members each, take several tens of PB of data per year of LHC operation, even after highly de-selective triggers [21]. As the machine is progressively upgraded, these annual data volumes will increase giving a total data sample between 10 and 100EB by the end of active operation, around 2035 – 2040 according to current plans. The re-use of this data within the collaborations during this period is fundamental, including the ability to reproduce past analyses. Recently, all four of the main LHC collaborations (ALICE, ATLAS, CMS and LHCb) have agreed open access policies [27] whereby significant sub-sets of the data are released to other scientists, as well as the general public, after embargo periods. Amongst other things, this period allows the data to be fully prepared for public consumption, along with the corresponding documentation and computing environment – to do this in pseudo real-time would be impossible with the resources that are available.

The Computing needs of the LHC experiments are met via the Worldwide LHC Computing Grid (WLCG) [34] that consists of a Tier0 site at CERN (with a remote facility also in Budapest), some ten Tier1 sites elsewhere in Europe, North America and Asia and some hundred Tier2 sites around the world. Whilst a full description of WLCG is outside the scope of this paper it is important to point out that the purpose of the Grid is for rapid processing and analysis of the data by scientists and it is optimized for such. Its use by the general public and / or for educational outreach is not compatible with the way it is designed, resourced or run. This has some implications for the re-use of preserved data and these aspects are described in detail below.

Through a combination of certification for the core preservation services together with additional metrics associated with these key Use Cases we are able to measure how close we are to achieving and sustaining our goals. These Use Cases also form the basis of a Business Model for ensuring sustained funding over a period of (at least) several decades. Finally, we highlight where our experience differs from "conventional wisdom" and in particular tensions between on-going use of preserved data for which it was originally intended versus "Open Access" style usage for educational outreach and other purposes.

1.1 HEP-Wide Data Preservation

In this paper we focus on the data preservation services that are offered by CERN with the LHC experiments as a key Use Case. The current phase of data preservation in HEP was kick-started by a study group that was initiated in late 2008 by DESY [6], resulting in a Blueprint report [23] in May 2012. The study group evolved to include all major HEP laboratories worldwide and reported to the International Committee for Future Accelerators (ICFA) [14], emphasizing its truly global nature.

The Study Group made the following observation:

"Data from high-energy physics (HEP) experiments are collected with significant financial and human effort and are mostly unique. An inter-experimental study group on HEP data preservation and long-term analysis was convened as a panel of the International Committee for Future Accelerators (ICFA). The group was formed by large collider-based experiments and investigated the technical and organisational aspects of HEP data preservation. An intermediate report was released in November 2009 addressing the general issues of data preservation in HEP. This paper includes and extends the intermediate report. It provides an analysis of the research case for data preservation and a detailed description of the various projects at experiment, laboratory and international levels. In addition, the paper provides a concrete proposal for an international organisation in charge of the data management and policies in high-energy physics."

The DPHEP study group identified the following priorities, in order of urgency:

- ***Priority 1: Experiment Level Projects in Data Preservation.*** *Large laboratories should define and establish data preservation projects in order to avoid catastrophic loss of data once major collaborations come to an end. The recent expertise gained during the last three years indicate that an extension of the computing effort within experiments with a person-power of the order of 2-3 FTEs leads to a significant improvement in the ability to move to a long-term data preservation phase. Such initiatives exist already or are being defined in the participating laboratories and are followed attentively by the study group.*
- ***Priority 2: International Organisation DPHEP.*** *The efforts are best exploited by a common organisation at the international level. The installation of this body, to be based on the existing ICFA study group, requires a Project Manager (1 FTE) to be employed as soon as possible. The effort is a joint request of the study group and could be assumed by rotation among the participating laboratories.*
- ***Priority 3: Common R&D projects.*** *Common requirements on data preservation are likely to evolve into inter-experimental R&D projects (three concrete examples are given above, each involving 1-2 dedicated FTE, across several laboratories). The projects will optimise the development effort and have the potential to improve the degree of standardisation in HEP computing in the longer term. Concrete requests will be formulated in common by the experiments to the funding agencies and the activity of these projects will be steered by the DPHEP organisation.*

These priorities could be enacted with a funding model implying synergies from the three regions (Europe, America, Asia) and strong connections with laboratories hosting the data samples.

1.2 Worldwide HEP Data Preservation

Since 2013, the former study group has evolved to a Collaboration and has built partnerships with data preservation efforts and projects in other disciplines. We have benefitted significantly from such partnerships and believe that it is key to offering long-term sustainable services. A Status Report summarizing the progress made since the publication of the Blueprint is available here [22].

The main messages contained in that report are as follows:

- Significant progress has been made in the past years regarding our understanding of, and implementation of services and solutions for, long-term data preservation for future re-use;
- **However, continued investment in data preservation is needed: without this the data will soon become unusable or indeed lost (as history has told us all too many times);**
- **Some of this investment can be done centrally, e.g. by providing bit preservation services for multiple experiments at a given laboratory, whilst important elements need to be addressed on an experiment-by-experiment basis.**
- Funding agencies – and indeed the general public – are now understanding the need for preservation and sharing of "data" (which typically includes significant metadata, software and "knowledge") with requirements on data management plans, preservation of data, reproducibility of results and sharing of data and results becoming increasingly important and in some cases mandatory;
- The "business case" for data preservation in scientific, educational and cultural as well as financial terms is increasingly well understood: funding beyond (or outside) the standard lifetime of projects is required to ensure this preservation;
- A well-established model for data preservation exists – the Open Archival Information System (OAIS). Whilst developed primarily in the Space Data Community, it has since been adopted by all most all disciplines – ranging from Science to Humanities and Digital Cultural Heritage – and provides useful terminology and guidance that has proven applicable also to HEP;
- **The main message – from Past and Present Circular Colliders to Future ones – is that it is never early to consider data preservation: early planning is likely to result in cost savings that may be significant. Furthermore, resources (and budget) beyond the data-taking lifetime of the projects must be foreseen from the beginning.**

1.3 DPHEP 2020 Vision

The "vision" for DPHEP – first presented to ICFA in February 2013 – a consists of the following key points:

- By 2020, all **archived data** – e.g. that described in DPHEP Blueprint, including LHC data – should be easily

[1] Electrons are believed to be fundamental, whereas nuclei are not, nor are their components.

[2] A simple analogy might be taking a photograph versus flipping a coin. If you take "the same" photograph tomorrow, or in a year's time, you are in fact recording something different. If you flip a coin ten times today, tomorrow, or in one year (statistically) it makes no difference.

findable and fully **usable** by the **designated communities** with clear (Open) access policies and possibilities to annotate further

- Best practices, tools and services should be well run-in, fully documented and sustainable; built in common with **other disciplines**, based on standards
- There should be a **DPHEP portal**, through which data / tools may be accessed
- **Clear targets & metrics** to measure the above should be agreed between **Funding Agencies, Service Providers** and the **Experiments (Collaborations).**

Although there is clearly much work still to be done, this vision looks both achievable and the timescale for realizing it has been significantly reduced through interactions with other (non-HEP) projects and communities. This is an important message for other projects and disciplines – collaboration can benefit us all.

2. BUSINESS CASE FOR PRESERVATION

Successful data preservation can only be performed if firstly one understands the motivation for such preservation – who will be the eventual re-users of the data, what is or will be the knowledge base of these re-users and what types of re-use are desired, for example for scientific, educational or simply cultural reasons. Secondly, it is clear that it will require resources and so the potential benefits, ideally in terms of a cost-benefit analysis, are desirable.

Following numerous discussions, a set of common Use Cases has been agreed across the 4 main LHC experiments. With some small provisos, these are also valid for other HEP experiments worldwide.

The basic Use Cases are as follows:

1. Bit preservation as a basic "service" on which higher level components can build;
   - Motivation: Data taken by the experiments should be preserved at least during the lifetime of the experiments and preferably until "redundant".
2. Preserve data, software, and know-how[3] in the collaborations;
   - This is the foundation for the long-term DP strategy
   - Analysis reproducibility: Data preservation alongside software evolution
3. Share data and associated software with (wider) scientific community, such as theorists or physicists not part of the original collaboration
   - This brings additional requirements:
     - Storage for the released data, distributed computing resources to access and process it
     - Accessibility issues, intellectual property
     - Formalising and simplifying data formats and analysis procedures
     - Documentation targeted at the specific consumer communities (lower knowledge base).
4. Open access to reduced data set to general public
   - Education and outreach
   - Continuous effort to provide meaningful examples and demonstrations

2.1 REQUIREMENTS FROM FUNDERS

Increasingly, Funding Agencies (FAs) are requiring Data Management Plans (DMPs) as part of the project approval process. Although these differ in detail from agency to agency, there is nevertheless significant commonality. Using the Guidelines for Data Management Plans in the European Union's Horizon 2020 programme as an example, DMPs should cover, at a minimum, the following:

A DMP describes the data management life cycle for all datasets to be collected, processed or generated by a research project. It must cover:

- *the handling of research data during & after the project*
- *what data will be collected, processed or generated*
- *what methodology & standards will be applied*
- *whether data will be shared / made open access & how*
- *how data will be curated & preserved*

More details are given regarding data sharing and preservation. Furthermore, other Funding Agencies stress reproducibility of results.

For a worldwide project such as the LHC, compliance with the requirements from multiple funding agencies is required. We thus refer to "an intelligent superset" of these requirements that includes not only those from the funders but also those needed internally within the project for its own scientific needs. In fact, we see remarkable synergy between the Use Cases presented above and this superset of requirements.

We believe that the information required to produce and maintain a DMP is typically available within a project. Presenting it in a common format is much more than a "contractual" requirement – it can – and should – be of use within the project, for data sharing and outreach portals, as well as to compare data management and preservation strategies across even heterogeneous communities.

The DMP for the LHC experiments is given further below.

2.2 COST MODELS

In order to estimate the cost for bit preservation over a long period of time we have developed a simple cost model that is freely available and can be readily adapted to the parameters of other projects. It is based on publicly available pricing information and technology predictions and – for the LHC – assumes an initial archive size with increasing growth rate, a 10% disk cache in from of the tape (robotic) store and regular (triennial) migration to new denser media. Whereas "conventional wisdom" is that the cost of storage will inevitably spiral out of control, at least for our predicted growth this does not appear to be the case. (Although the data rates do increase with time, the detectors are essentially "static" and not, for example, doubling in the number of channels every 18 months or so).

[3] Additional Use Cases help to define whether the "know-how" has been adequately captured. See the Analysis Preservation section for further details.

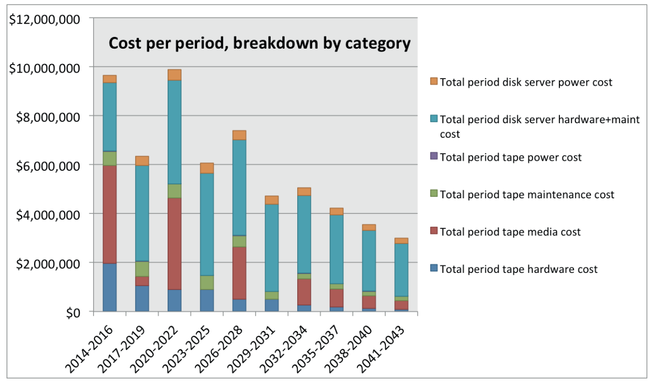


Figure 1 - Bit Preservation Costs per Period

Here we see that the costs decrease with time despite a significant increase in the total data stored. Of course, it is possible to construct data growths that will exceed the benefits from increase in storage density but at least for one major project this does not seem to be the case.

Naturally, there are large uncertainties in such a model – will tape density continue to increase as foreseen, is there sufficient market demand and many other factors.

These numbers are nonetheless far from negligible – how do they compare with the overall cost of computing and / or the cost of the LHC project as a whole and what are the benefits of this bit preservation?

2.3 PRESERVATION BUSINESS CASE

The cost of bit preservation – certainly only one element of an overall data preservation strategy – can be seen to be small compared to the overall costs of computing for such a project as the LHC and smaller still when compared to the entire project (the collider, the detectors, the collaborations and so forth). On the other hand, the benefits can be measured quasi-directly: in terms of the scientific publications and PhDs that it enables, as well as indirectly in terms of technological, educational and cultural spin-offs. What is observed by all HEP experiments is that analysis of the data, publications and presentations continue many years after the end of data taking and their use for educational outreach continues even further. Again, we are "saved" by the fact that we make measurements and not observations – the duration for which the data (and associated software, metadata, documentation and so forth) should be maintained is perhaps a few decades – for example until a new and more powerful facility is available and first comparisons have been made – and not "forever" as is desirable with observations. (In fact, as the costs for bit preservation tend to zero there is no argument to delete the data even after this period and successful "data resurrection" has been achieved in the past – it requires only sufficient scientific motivation, such as a new or improved theoretical model and/or the discovery of a new particle that should have been visible in the old data).

3. DATA MANAGEMENT AND PRESERVATION PLAN FOR WLCG

Based on the Horizon 2020 guidelines [10], a Data Management Plan[4] for the 4 main LHC experiments – in the context of the Worldwide LHC Computing Grid – has been prepared. The following subsections specify this plan, together with the associated guidelines *quoted verbatim in italics*.

3.1 Data set reference and name

Identifier for the data set to be produced.

This Data Management Plan (DMP) refers to the data set generated by the 4 main experiments (also know as "Collaborations") currently taking data at CERN's Large Hadron Collider (LHC).

These experiments are ALICE, ATLAS, CMS and LHCb. For the purpose of this plan, we refer to this data set as "The LHC Data".

In terms of Data Preservation, the software, its environment and associated documentation must also be preserved (see below).

Further details can be found at the DPHEP portal site, described further below.

3.2 Data set description

Description of the data that will be generated or collected, its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.

The 4 experiments referenced above have clear scientific goals as described in their Technical Proposals and via their Websites. These are accessible through the official catalogue of all CERN experiments that is maintained by the CERN Research Board, the CERN Grey Book [25].

Hundreds of scientific publications are produced annually.

The data is either collected by the massive detectors of the above experiments (the raw data), is derived from it, or is the result of the simulation of physics processes according to theoretical models and the simulated response of the detector to these models.

Similar data – but at lower energies – have been produced by previous experiments and comparisons of results from past, present and indeed future experiments is routine.

The data behind plots in publications is made available since many decades via an online database, HEPData, described below.

Re-use of the data is made by theorists, by the collaborations themselves, by scientists in the wider context as well as for Education and Outreach.

3.3 Standards and metadata

Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.

The 4 main LHC experiments work closely together through the WLCG Collaboration on data management (and other) tools and applications. At least a number of these have found use outside the HEP community but their initial development has largely been driven by the scale and timeline of the above. The ROOT framework [20], in particular, is used as "I/O library" (and much more) but all LHC experiments and is a *de-facto* standard within HEP, also across numerous other laboratories.

The meta-data catalogues are typically experiment-specific although globally similar. The "open data release" policies foresee the available of the necessary metadata and other "knowledge" to make the data usable (see below).

3.4 Data sharing

Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups.

[4] Much of this work is required as part of the formal approval process of an experiment and / or as part of data sharing and preservation planning.

Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.).

In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).

The 4 LHC experiments have policies for making data available, including reasonable embargo periods, together with the provision of the necessary software, documentation and other tools for re-use.

Data releases through the CERN Open Data Portal (see below) are published with accompanying software and documentation. A dedicated education section provides access to tailored datasets for self-supported study or use in classrooms. All materials are shared with Open Science licenses (e.g. CC0 or CC-BY) to enable others to build on the results of these experiments. All materials are also assigned a persistent identifier and come with citation recommendations.

3.5 Archiving and preservation

Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.

The long-term preservation of LHC data is the responsibility of the Tier0 and Tier1 sites that form part of the WLCG Collaboration. A Memorandum of Understanding [33] outlines the responsibilities of sites that form part of this collaboration (Tier0, Tier1s and Tier2s).

In the case of the Tier0 and Tier1s, this includes “curation” of the data with at least two copies of the data maintained worldwide (typically 1 copy at CERN and at least 1 other copy distributed over the Tier1 sites for that experiment).

The costs for data storage and “bit preservation” form part of the resource requests that are made regularly to the funding agencies. A simply cost model shows that the annual storage costs – even including the anticipated growth – go down with time and remain within the funding envelope foreseen. (The integrated costs of course rise).

Personnel from the Tier0 and Tier1 sites have followed training in ISO 16363 certification – A Standard for Trusted Digital Repositories – and self-certification of these sites is underway.

Any data generated on external resources, e.g. Clouds, is copied back for long-term storage to the Tier0 or Tier1 sites. The eventual long-term storage / preservation of data in the Cloud would require not only that such services are cost effective but also that they are certified according to agreed standards, such as ISO 16363.

The data themselves should be preserved for a number of decades – at least during the active data taking and analysis period of the LHC machine and preferably until such a time as a future machine is operational and results from it have been compared with those from the LHC.

The total data volume – currently of the order of 100PB – is expected to eventually reach 10-100 EB (in circa 2035 – 2040).

Additional services are required for the long-term preservation of documentation (digital libraries), the software to process and/or analyse the data, as well as the environment needed to run these software packages.

All such services are the subject of the on-going self-certification.

4. SERVICE PORTFOLIO

In this section we address the high level Use Cases in the order presented, namely:

1. Bit Preservation;
2. Preserving data, software and know-how within the “producer collaborations”;
3. Share data and associated software with (larger) scientific community;
4. Open access to reduced data set to general public.

One key concern is the ability to “reproduce” physics analyses published in the past. The scientific data underlying publications by the CERN experiments is complex and requires a rich ecosystem of descriptive information. The data published in scientific papers is produced from derived data using software specific to that analysis. That software, a reference to and provenance of the derived data used, the computing environment, and the analysis documentation is catalogued in the CERN Analysis Preservation Portal. Some derived data is catalogued with the necessary access protocols on the CERN Open Data Portal. The information on the CERN Open Data Portal allows a virtual machine to access data on the CERN bit preservation infrastructure. These issues are discussed in more detail below.

4.1 Bit Preservation

We predict a total data volume of a few exabytes (EB) from the LHC experiments by the 2030s [18] and a final data volume between 10 and 100 EB. In fact, Influx rates from CERN’s Large Hadron Collider experiment are expected to augment from currently 40 Petabytes / year to around 600 Petabytes / year in a few years time, therefore reaching archive volumes at the Exabyte-scale. The data from the rest of the CERN experimental programme can be archived easily alongside this massive LHC dataset. At this scale, bit-rot – the tendency of content in storage to become corrupt over time – becomes unavoidable. Reasons for bit rot are multiple, the most important ones being: wear out and breakage of media components (such as disk drive head crashes, snapping tapes, RAID controller failures); undetected bit flips during data transmission; hardware or media obsolescence; or environmental hazards (fire, water leaks, dust contamination).

- In order to minimise data loss and improve archive reliability, CERN has implemented storage verification and preservation services on top of its tape-based Mass Storage System (CASTOR [2]). These measures include notably:
- Regular Media verification: Every time a tape is written to, it will be subject to a verification process that consists in checking critical areas, namely the first and last ten files of the tape, as well as 10 random files across the tape, and validating the metadata (such as file sizes and checksums). When a tape is filled, all its contents will be verified. In addition, all tapes are re-verified approximately every 2 years, ensuring also the correctness of older repository data.
- Controlled media lifecycle: Media at CERN is typically kept in production for not longer than two drive generations (typically 6-8 years). This is well below the physical media lifetime, which is around 20-30 years. While tape media is well-known for its longevity (30 years or more), the associated hardware infrastructure typically enters obsolescence after 4-6 years, after which it becomes difficult to find replacements for tape drives, firmware patches or software drivers for new operating system versions. In addition, newer media usually comes with increased reliability over older generations. Last but not least, by migrating existing data to newer-generation and higher-capacity media, less cartridges will be required and expenses in additional tape libraries and floor space can be avoided.
- Reducing tape mounts: In order to reduce media wear out and to increase efficiency, a policy-driven engine examines each tape read request and decides on whether to grant a tape mount or postpone it. This takes into account criteria such as user/group priority, number of files and amount of volume to be read, waiting time, and concurrent drive usage by the user/group. Since deployment in 2010, and despite continuous file and volume recall increases, the average number of daily tape read mounts has been reduced from over 4000/day to 1500/day.
- Data redundancy: For smaller communities, such as the former LEP experiments, secondary file copies can be created. These second data copies are stored in a separate library residing in a different physical building.
- Protecting the physical link: In order to increase the reliability of data transfers between the disk caches and tape servers, CERN has implemented support for SCSI Logical Block Protection [19]. This mechanism protects the path between the data source and the tape media (e.g. FC interface and physical link, internal drive data channels, etc.) against errors such as link-level bit flips. It works by pre-calculating and appending a CRC code for each data block sent to the tape drive, which is then re-calculated and verified at every transfer back and forth to tape.
- Protecting the operating environment: Tape media is vulnerable to contamination from airborne dust particles that can land on the rollers, reels or heads. These can cause scratches on the tape as it is being mounted or wound on the tape drive. With tape media bit sizes smaller than a bacterium or the particles emitted by a car exhaust, any damage to the tape can destroy significant amounts of data. CERN has prototyped and built custom environmental sensors that are hosted in the tape libraries, sampling the same airflow as the surrounding drives [8]. The sensor continuously samples the surrounding air and issues alarms if airborne particle density, humidity or temperature crosses configurable thresholds.

These measures have helped reducing the number of annual file losses by two orders of magnitude. For the period 2012-2015, the annual bit loss rate is in the order of $5*10^{-16}$. This rate can still be improved as it is still three to four orders of magnitude above the undetected bit error rate for enterprise-level tape drives, which can be considered as the upper ceiling to reach in terms of reliability.

4.2 Software Preservation

The HEP community has a long tradition of sharing and developing common, open-source software stacks within international collaborations. The software systems required to operate on LHC data comprise physics simulation and reconstruction algorithms to determine physics processes from detector signals, data analysis frameworks to extract (new) scientific knowledge from data sets, and distributed systems for data access and compute job management. Altogether, HEP software stacks add up to tens of millions of lines of code, half a dozen different languages and tens to hundreds of modules with dependencies on each other. Several millions of lines of code are specific to an experiment and there are numerous dependencies on standard software, most notably the GNU/Linux operating system and language compilers and interpreters. The support life cycle of the 3rd party software components is much shorter than the envisaged preservation period of several decades. Operating system versions are supported for a maximum of 5-10 years, for instance, and most developers abandon their software releases much earlier.

Running experiments invest many person months in the porting and the validation of their software stacks, coordinated by a dedicated software librarian. For a decommissioned experiment, that is one that is no longer in its data-taking phase, such an amount of effort would be impractical. Hardware virtualization (such as KVM [17], Xen [35] and VirtualBox [36]) and container virtualization (such as Docker [7]) provide a potential solution. Virtualization allows for execution of a frozen, historic software environment on contemporary hardware and operating systems. In a straightforward application of virtualization technology, a software environment is frozen in the form of a disk image, a large and opaque stream of bytes containing all the necessary software binaries. This approach tends to be clumsy and too rigid for HEP software. In order to be useful, even “frozen” environments need to stay open for minor modifications: patches to faulty algorithms, updates to physics models, updated tuning parameters for simulation algorithms, new configuration for data access software and so on. Software development communities have long solved similar problems by version control systems. Version control systems only store a track of changes to the source code and at the same time they can provide access to the state of a directory tree at any given point in the past.

In an attempt to get similar benefits for compiled and configured software stacks, since several years HEP experiments install all released software components in its final configuration on the versioning, open source, and distributed file system CernVM-FS [5]. By selecting different versions in the history provided by the file system, experiments can access any software state ever released. Thus we can separate virtualization – in this case handled by CernVM [26] – itself from the concerns of accessing software binaries. A minimal and stable virtual machine or container (~20MB) connects to the remote file system CernVM-FS that hosts the operating system and software binaries. By selecting different states of the versioned file system, experiments can go back and forth in time and create software environments compatible with Red Hat Enterprise Linux (RHEL) 4 to RHEL 7 (spanning 15+ years) with the very same virtual machine on the very same hardware. Concretely, we have demonstrated that by resurrecting the software of the ALEPH experiment at LEP more than 15 years[5] after the experiment was decommissioned. Contemporary virtual machines provide data access tools and middleware with support for the latest network protocols and security settings. Containers inside the virtual machines spawn historic operating system and application software environments. Data is provided from the container host to the historic applications through the very stable POSIX file system interface.

Among the currently active HEP experiments, many operate dedicated CernVM-FS services to satisfy their day-to-day needs for global software distribution. These services are operated in an “append-only” mode, so that software versions, once released to the production infrastructure, remain readily available for future use. Due to the file system’s internal data de-duplication, this model proved to be sustainable even for the largest users. After more than five years of experience with LHC experiment software and more than hundred million

[5] Data continue to have scientific value and underpin publications and even PhDs long after data taking has ended. Porting and validating the associated software as well as handling changes in data formats is a story in itself and outside the scope of this paper.

registered files (software only!), the storage volume is still at only a few terabytes.

4.3 HEPData

HEPData [11] did not originate as a CERN service but deserves a specific mention as it has provided access to data behind physics publications for several decades. Founded and run by Durham University, it "*has been built up over the past four decades as a unique open-access repository for scattering data from experimental particle physics. It currently comprises the data points from plots and tables related to several thousand publications including those from the Large Hadron Collider (LHC)".*

Thus it is complementary to the portals offered by CERN and a transition will soon be made to a new HEPData site, hepdata.net, based on Invenio [15], developed in collaboration with INSPIRE [13].

4.4 Analysis Preservation Portal

Research outputs in physics range from data, software and documentation to the "traditional" publication – while so far only the latter is preserved and published openly. Currently, the user-generated content is scattered around various tools and services within the individual experiments and it is difficult to relate the individual elements to a specific analysis and research result. To enable others to build on collaborators' findings and to foster reproducible research it is important to preserve and provide (internal) access to the wider range of related and supplementary materials.

Hence, a new service is developed in close collaboration with the research community, i.e. the LHC collaborations (ALICE, ATLAS, CMS, LHCb). "CERN Analysis Preservation" is under development to capture research objects while researchers are in the process of conducting their research. The new service is being built on the latest release of Invenio, an open source digital library software developed at CERN and addresses the complex but essential Use Cases for Analysis Preservation [1]. For example:

- An analysis that is underway has to be handed over, e.g. as someone is leaving the collaboration;
- A previous analysis has to be repeated;
- Data from different experiments have to be combined.

The service aims at preserving the insider knowledge about a physics analysis. The researchers provide metadata about the data, software, configurations options, high-level physics information, documentation, instructions, links to presentations, quality protocols, internal notes, etc. The system connects to selected databases in the experiments so that information can be retrieved automatically, therefore information is up to date and researchers do not have to spend much extra effort on using this service. The service also facilitates versioning to accommodate the work in progress nature of the materials.

Once the service reaches production level, it will allow users of the collaboration for the first time to search for related user generated content they might be interested in. Furthermore, it is expected to allow internal analysis tools to plug into this new service to make use of the central information resource.

It should be noted that CERN Analysis Preservation is a "closed access" service, as it deals with the early stages of the data analysis process before the results are submitted for publications. The access is thus restricted to the individual LHC collaboration. After an analysis is approved for publication the CERN Analysis Preservation service may (upon request by the researcher) push parts of information to the public CERN Open Data portal and the INSPIRE services. Hence, in combination with the other preservation services at CERN, CERN Analysis Preservation should help fostering preservation and Open Science practices in the community.

4.5 Open Data Portal

Corresponding to the LHC data policies [27], a service was needed to serve large scale and complex datasets, together with underlying virtual analysis environment, an example software code, and supporting materials. Existing services, such as WLCG were not (by construction) suited to accommodate the needs for the sharing of complex and big datasets. Hence, the public CERN Open Data Portal [4] was launched in November 2014, providing data and accompanying software and tools for education and research purposes to the public. To give an example: the annual CMS data release of data from 2010 focused on primary and derived data, which amount to a volume of 27TB; the 2011 release comprised simulated data, detailed information about configuration and triggers and, hence, resulted in several hundred terabytes. All LHC collaborations have already shared data through this service.

Special emphasis was given on providing comprehensive metadata for the objects that are shared and on an appealing user interface. To serve the target groups, the physicists and the non-physicists, best the repository was presented with a modern website layout. A close collaboration of CERN IT, the Scientific Information Service, and the physics collaborations ensured that sufficient (interactive) tools, accompanying documentation and metadata are provided in an understandable way to facilitate future reuse. Following best practices, materials shared through this services are issued a persistent identifier so that reuse (i.e. citations) can be tracked on INSPIRE.

CERN Open Data Portal and INSPIRE are based on the Invenio digital library software.

4.6 DPHEP Portal

The portal of the DPHEP Collaboration [29] provides a human readable reference of sites and services used by the worldwide effort to preserve HEP data. The aim is to make the HEP data findable, accessible, interoperable, and re-usable [30]. The landing page lists the member institutes with their preferred data preservation portal. Members of the collaboration provide detailed information on current status of their data regarding: bit preservation, data, documentation, software, use cases, target audience, value, uniqueness, resources, issues, and outlook. The portal makes the applicable preservation and access policies available and provides relevant contact details. Agreeing on a common way of presenting the status at the different sites and laboratories took several years of elapsed time and helps to highlight commonalities and areas for shared developments, services and / or tools.

Furthermore, the portal provides a central reference point to meetings, documents, and services organizing the HEP data preservation effort.

Institutes / organisations referenced from the portal include Brookhaven National Laboratory, Fermi National Laboratory and Stanford Linear Accelerator Laboratory, all in the US; CERN, CSC, DESY, INFN, IN2P3 and STFC in Europe as well as IPP in Canada, IHEP in China and KEK in Japan.

5. CERTIFICATION OF REPOSITORIES

It is widely accepted that certification of repositories is at least a best practice as part of a long-term data preservation strategy. Whilst there are a number of certification frameworks in use [12], that covered by ISO 16363 [16], based on the OAIS reference model [31], is considered to be the most comprehensive and even ambitious. Moreover, it matches our current practices more closely than other such frameworks.

In the context of the LHC experiments, the "repository" for long-term storage consists of the WLCG Tier0 site (CERN) plus the Tier1 sites spread around the world. A copy of all "archive" data is maintained at CERN with at least one additional copy being spread over the Tier1 sites that serve that experiment.

Representatives of these sites have undergone training in the application of ISO 16363 and self-certification is underway with a goal of covering at least the WLCG Tier0 prior to iPRES 2016. This would be a first step – formalizing some of the issues in terms of official policies will not be achieved on this time frame. Similarly, including all of the Tier1 sites will take some additional time, as would extending this certification to cover all experiments whose data is archived at CERN and/or all projects supported by the Tier1s [37], many of which are multi-disciplinary.

We believe that this will greatly enhance the transparency and long-term sustainability of our overall long-term data preservation strategy. In particular, it will formalize and institutionalize many of the practices, services and strategies described in this paper. Furthermore, any gaps identified will help us improve our strategy for the future.

6. CONCLUSIONS

We have presented the services offered by CERN for the long-term preservation of the data from the LHC experiments, along with a business case and a draft Data Management Plan. Key to this DMP is the on-going self-certification of the archive sites according to the ISO 16363 standard – the most rigorous of the various certification standards currently available. Goals for data sharing and reproducibility of results have been shown and by constantly monitoring these we are able to measure we are meeting are targets. Whilst ISO 16363 is discipline agnostic, at least some details of our requirements and practices for sharing and reproducibility require domain-specific knowledge. Furthermore, our experience with data sharing and re-use is still relatively young: as more data is released, crossing the PB threshold and well beyond, new issues will arise and fresh lessons will be learned. However, we strongly believe that the more these issues are addressed in common the more everyone benefits. We have also shown the need for separate (but linked) infrastructures for different purposes: WLCG provides the main processing, analysis and archival facilities whilst the Portals perform tasks related to reproducibility, data sharing and outreach. Such a separation is particularly important during the active data taking stage of an experiment and may become less critical with time but we believe that it should not be overlooked. Finally, our cost model shows that, at least for bit preservation, and assuming no major technological surprises, our overall budget is sustainable in both medium and long term.

7. ACKNOWLEDGMENTS

Although this paper focuses on the services offered by CERN, the work is not done in isolation and owes much not only to the DPHEP Collaboration [28] but also to many others active in the data preservation world, including members of the Alliance for Permanent Access (APA) [24], the Research Data Alliance (RDA) [32] and its many Working and Interest Groups as well as numerous data preservation projects, workshops and conferences worldwide.

Furthermore, it is the experiments themselves that not only drive the requirements described above but also need to do a significant amount of work to prepare, clean, document and share their data as well as support access to it – these services are for them, as well as potential re-users of the data.

8. REFERENCES

[1] CAP Use Cases: http://dx.doi.org/10.5281/zenodo.33693.

[2] CASTOR homepage http://cern.ch/castor.

[3] CERN – the European Organisation for Nuclear Research – see http://home.cern/.

[4] CERN Open Data Portal – see http://opendata.cern.ch/.

[5] CernVM-FS – see https://cernvm.cern.ch/portal/filesystem.

[6] Data Preservation in High Energy Physics (DPHEP): see http://www.dphep.org/.

[7] Docker – see https://www.docker.com/.

[8] Dust sensors: Data Centre Environmental Sensor http://www.ohwr.org/projects/dces-dtrhf-ser1ch-v1/wiki.

[9] European Strategy for Particle Physics – see https://indico.cern.ch/event/244974/page/1736-welcome.

[10] Guidelines for Data Management Plans in Horizon 2020: see http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

[11] HEPData – see http://hepdata.cedar.ac.uk/; https://hepdata.net/.

[12] http://www.trusteddigitalrepository.eu/Memorandum%20of%20Understanding.html.

[13] INSPIRE: the HEP information system – see http://inspirehep.net/.

[14] International Committee for Future Accelerator (ICFA) – see http://icfa.fnal.gov/.

[15] Invenio – see http://invenio-software.org/.

[16] ISO 16363: Audit and Certification of trustworthy digital repositories – see http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510.

[17] KVM – a Kernel-based Virtual Machine – see http://www.linux-kvm.org/page/Main_Page.

[33] The WLCG Memorandum of Understanding – see http://wlcg.web.cern.ch/collaboration/mou.

[34] The Worldwide LHC Computing Grid (WLCG) – see http://wlcg.web.cern.ch/.

[35] The Xen project – see http://www.xenproject.org/.

[36] VirtualBox – see https://www.virtualbox.org/.

[37] WLCG Tier1 sites: see http://wlcg-public.web.cern.ch/tier-centres.

[18] LHC data volume predictions: Bird I et al. 2014 Update of the Computing Models of the WLCG and the LHC Experiments *CERN-LHCC-2014-014*.

[19] Logical block protection: Butt K 2007 Tape end-to-end data protection proposal http://www.t10.org/ftp/t10/document.07/07-374r0.pdf.

[20] ROOT – a modular scientific software framework – see https://root.cern.ch/.

[21] See http://home.cern/about/computing/processing-what-record for further details.

[22] Status Report of the DPHEP Collaboration, February 2016: https://zenodo.org/record/46158#.Vvo8zBL5jBI.

[23] Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics: http://arxiv.org/pdf/1205.4667.

[24] The Alliance for Permanent Access (APA) – see http://www.alliancepermanentaccess.org/.

[25] The CERN database of experiments – "the Grey Book": see https://greybook.cern.ch/greybook/.

[26] The CernVM Software Appliance – see https://cernvm.cern.ch/portal/filesystem.

[27] The data (access) policies of the four LHC experiments are available here: http://opendata.cern.ch/collection/Data-Policies.

[28] The DPHEP Collaboration – see https://hep-project-dphep-portal.web.cern.ch/.

[29] The DPHEP Portal http://dphep.web.cern.ch.

[30] The FAIRport initiative http://www.datafairport.org.

[31] The Open archival information system (OAIS) -- Reference model – see http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284.

[32] The Research Data Alliance (RDA) – see https://rd-alliance.org/node

Designing Scalable Cyberinfrastructure for Metadata Extraction in Billion-Record Archives

Gregory Jansen
University of Maryland
jansen@umd.edu

Richard Marciano
University of Maryland
marciano@umd.edu

Smruti Padhy
NCSA/UIUC
spadhy@illinois.edu

Kenton McHenry
NCSA/UIUC
mchenry@illinois.edu

ABSTRACT

We present a model and testbed for a curation and preservation infrastructure, "Brown Dog", that applies to heterogeneous and legacy data formats. "Brown Dog" is funded through a National Science Foundation DIBBs grant (Data Infrastructure Building Blocks) and is a partnership between the National Center for Supercomputing Applications at the University of Illinois and the College of Information Studies at the University of Maryland at College Park. In this paper we design and validate a "computational archives" model that uses the Brown Dog data services framework to orchestrate data enrichment activities at petabyte scale on a 100 million archival record collection. We show how this data services framework can provide customizable workflows through a single point of software integration. We also show how Brown Dog makes it straightforward for organizations to contribute new and legacy data extraction tools that will become part of their archival workflows, and those of the larger community of Brown Dog users. We illustrate one such data extraction tool, a file characterization utility called Siegfried, from development as an extractor, through to its use on archival data.

Keywords

Computational archival science, Digital curation, Data mining, Metadata extraction, File format conversion, Brown Dog, Cyberinfrastructure, Big data

1. INTRODUCTION

1.1 The Data Observatory in Maryland

The Digital Curation Innovation Center (DCIC) at the UMD College of Information Studies ("Maryland's iSchool") is building a 100 Million-file data observatory (called CI-BER – "cyberinfrastructure for billions of electronic records") to analyze big record sets, provide training datasets, and teach students practical digital curation skills. At 100 Million files we seek to anticipate the billion-file scale, testing approaches on collections one order of magnitude removed. The DCIC is contributing to a $10.5M National Science Foundation / Data Infrastructure Building Blocks (DIBBs)-funded project called "Brown Dog", with partners at the University of Illinois NCSA Supercomputing Center. The DCIC is also partnering with industry storage leader NetApp and an archival storage startup company, Archive Analytics Solutions, Ltd. As a newly formed center for digital curation, we are fortunate to collaborate on a project that addresses large-scale challenges and has extraordinary strategic potential. The Brown Dog project[1] is the largest of the implementation awards to date under the NSF Data Infrastructure Building Blocks (DIBBs) program. Brown Dog is creating web-scale infrastructure services that open up data collections for appraisal, analysis, and reuse. The approach has been described as creating a new infrastructure service for the web, like a domain name service (DNS) for data, the idea being that data-focused services are a missing component of the web we have today. The role of the DCIC in Brown Dog is to use these infrastructure services to enrich the collections and meet the curation challenges we face in data-driven research.

1.2 Digital Legacies and Format Debt

Heterogeneous data accumulate in research and memory institutions, but often remain locked in native formats and are not easily accessed or understood as rich, informative sources. Legacy files will often not open in current software readers or viewers. Moreover, their internal information remains opaque to modern search and analytic approaches. As the files accumulate, so do the missed opportunities to effectively exploit them. We refer to this accumulation of effectively opaque file formats as a type of institutional debt, "format debt", which we would quantify as the theoretical technology investment required to reveal the complete intellectual content of all the accumulated files.

The existence of a functionally opaque format is rarely due to the lack of available software. Many legacy and current software tools can process legacy file formats and reveal their intellectual content. From commercial Windows applications, such as CorelDraw, to Linux-hosted computer vision tools for image processing; the available software list goes on and on. The challenge with "format debt" is not the lack of software, but the instrumentation of all the associated software in a workflow.

Each software executes in a particular technical environment, including the required operating system and machine architecture. A technical expert must set up each software environment, devise a way of passing in a file, running the software, and interpreting the results. The myriad of old

[1] http://browndog.ncsa.illinois.edu/

and new software tools produce diverse output formats that rarely conform to current standards like JSON or RDF. These are the real barriers to instrumentation. It requires a significant investment to add each different software to the workflow. Unlike a digital collection of a single format, big archives of born digital materials contain thousands of formats. Big archives require a new strategy to tackle spiraling “format debt” and for that reason we explore integrations between archival collections and Brown Dog services.

1.3 CI-BER Testbed

Our testbed explores how the Brown Dog services [1, 7] can be applied within a large organization’s archives, to reveal the data within the diverse file formats of archival collections. We present a model architecture for a born-digital repository that inserts Brown Dog services into a repository workflow that also includes a scalable mix of search and analysis services, namely Indigo (Cassandra), Elasticsearch, and Kibana.

Several extractor tools have been developed for our testbed, with archives in mind. We use these as examples of community-developed tools added to the Brown Dog tools catalog, which is designed for such user contributions.

Lastly, the 100 Million files in the CI-BER data set are being used to systematically test the Brown Dog service APIs. These tests include load tests, to ensure that performance does not degrade under web-scale load, and qualitative tests of the services’ response to diverse file formats.

2. BACKGROUND

Brown Dog (BD) [1, 7] is a set of extensible and distributed data transformation services, specifically, data format conversions, named Data Access Proxy (DAP), and metadata extraction from data content, named Data Tilling Service (DTS). With ever increasing varieties of data formats, data sometimes becomes inaccessible due to obsolete software/file formats. The DAP, through a set of REST APIs, allows users to convert inaccessible data to accessible formats, thus unlocking valuation information. Similarly DTS, through a set of REST APIs, allows users to extract metadata, signatures, tags or any other possible feature information from a file’s content. Using the extracted information, files can be indexed and retrieved based on data content.

The scale and scope of the Brown Dog data service will often prompt comparisons with the SCAPE project for Scalable Preservation Environments[2]. Both are aimed at preserving data that resides in diverse file formats, but they are highly complementary. Brown Dog specifically focuses on building a cloud-based service, allowing data to broadly transcend format. In contrast SCAPE pursued diverse strategies, tools, and policies for digital preservation. SCAPE policies can provide a decision-making framework for ongoing preservation activities, whereas Brown Dog can provide the supporting metadata and format conversions. Brown Dog’s DAP and DTS REST services are a natural fit for use in SCAPE preservation work flows.

The DAP, built on top of Polyglot framework [4, 5], does file format conversions, i.e. it converts an input file to a given output format. It encompasses several Software Servers (SS). A SS uses a wrapper script (alternatively, known as *Converter* within BD) which wraps any piece of code, third party software, or library, to provide access to their conversion capabilities (e.g. open/save/convert) through a consistent REST interface. The wrapper script also provides information on the input and output formats supported by the underlying software, and thus, available through the SS. The DAP performs a format conversion by obtaining the available input/output formats from all the different SS, chaining these together for all possible conversion paths. It then finds the shortest conversion path from a given input format to a given output format. Lastly, the DAP performs the format conversion by passing the file data through the chain of SS along the shortest path. A user can write a wrapper script (or a custom converter) for the software she uses and contribute that to the Tools Catalog. Then a Software Server (SS) containing her script can be deployed within BD services and can be made available for other users to leverage.

The DTS is built on top of the Clowder framework [3, 6] and performs metadata extractions on-demand from a given input file’s content. The extraction process is triggered based on file mime type and then carried out by any appropriate extractors that are available. An extractor is a software process that listens for extraction requests from DTS on a message queue (RabbitMQ). It performs extraction of metadata from the file through analysis of the content, generating rich JSON-LD[3] metadata and tags, creating previews, and breaking files out into sections with more details. Each extractor then uploads the new information to Clowder, where it is combined with results from other extractors and made available through the DTS REST API. Using the pyClowder[4] library, a DTS user can write her own extractor that can use any piece of code, software, library, or webservice extraction functionality under the hood, and can potentially be deployed as a BD service for other users.

The DAP’s SS and the DTS’s extractors reside in a distributed environment such as the cloud. To handle heavy load, adapt to peak user demand and support heterogeneous architectures, another module, named the Elasticity Module (EM), has been incorporated into the BD. EM automatically scales up or down the DAP’s SS and DTS’s extractors based on user demands and loads. It monitors conversion and extraction requests in the RabbitMQ queues for SS and extractors. If any queue length exceeds a particular threshold, it launches another instance of that specific extractor or SS listening to the respective queue. Current implementation of EM uses the NCSA OpenNebula Openstack cloud for launching a new VM with a SS or extractor. It also has the option of using Docker as another layer of virtualization. Thus, this EM design allows BD services to dynamically grow/shrink based on demand and also ensures scalability. A detailed description of the BD services architecture can be found in [7].

3. BROWN DOG WITHIN THE CI-BER DATA INFRASTRUCTURE

The DCIC has accumulated a large-scale archival data repository in collaboration with the National Archives and Records Administration (NARA) consisting primarily of federal and community-sourced digital archives, both born-digital and digitized, which were part of an earlier NSF-funded CI-BER project [2]. The DCIC’s Digital Archives Repository for Research and Analytics (DARRA) houses more than 100 Million files and 72 Terabytes of unique, heterogeneous data.

The DCIC staff rely on assistance from the Division of Information Technology (Div IT) staff on campus and our industry partner for the Indigo repository software, Archival Analytics, to maintain the DARRA facility. DARRA equipment occupies half of a rack in a campus data center. Our storage array is maintained there on-site by NetApp services. The four virtual machine hosts are administered by Div IT staff. These relationships allows us to build and maintain the facility and carry out Brown Dog research in virtual machines, with a single software architect on staff. For production operations of this kind a dedicated system administrator is also required, providing for more formal change management, reporting, and vacation coverage.

The DCIC approach to curation infrastructure relies upon distributed databases, messaging, and virtual machine technology to eliminate bottlenecks and create linear scalability. The archival repository and related services are run on a cluster of four physical servers. The servers have high bandwidth connections to a peta-scale NetApp storage array. These physical servers play host to a constellation of guest virtual machines that run all the software. The DCIC is working with industry partners NetApp and Archive Analytics, Ltd., a big data startup, to build a scalable storage facility. Our catalog uses Archive Analytics’ repository software, Indigo; a resilient and scalable solution for storage virtualization and workflow automation. Based on the Apache Cassandra distributed database[5], Indigo gives us high performance data access over a standard cloud storage API (Cloud Data Management Interface – CDMI), which is critical to data processing activities. The Indigo repository software is to become a community open source initiative.

The Brown Dog service is integrated with the catalog through its two main web endpoints. The Data Access Proxy (DAP) exposes an API for arbitrary, on-demand file format conversion. The Data Tilling Service (DTS) provides an API that runs all of the metadata extraction tools that are appropriate to a submitted file. In our workflow the DAP and DTS APIs are called by very simple worker scripts that are written in Python. The worker scripts submit CI-BER data to the Brown Dog APIs and place the resulting metadata back into the Indigo catalog. A pool of Python workers are always available to handle this work, which is queued and tracked in a local RabbitMQ message queue. The Brown Dog workflow may be triggered automatically by placing a new file into Indigo, or it may be run on-demand, when the existing repository hierarchy is traversed.

For those building systems on a similar scale, the hardware in the data center rack totals 166,000 US dollars, which breaks down into $29,000 for the four servers and $137,000 for the NetApp storage array. The raw storage costs are a little over $190 per terabyte. The NetApp storage is used for the archives and also parceled out for other virtual machine needs, such as databases and index space.

4. CONTRIBUTE YOUR TOOL TO BROWN DOG AND SHARE

Researchers often build new tools for their research, in order to extract useful information from unstructured or semi-structured data and to do necessary file format conversions in the process. A lot of effort goes into developing such tools and such efforts are often unacknowledged. In addition, similar tool development efforts are repeated by multiple researchers within the same domain of research. Towards acknowledging such tool development efforts, the BD Tools Catalog (TC) was designed and implemented to be a framework for these contributions. BD TC is a web application where a user can upload any existing/new tool with conversion or extraction capabilities, e.g., imagemagick[6], and tesseract[7] and can share it with the research community. It has a web interface to upload BD tools (alternatively, known as BD scripts). A BD tool/script is a script that wraps the original software developed by a researcher, or third-party software, and make it deployable within the BD service. The TC can also deploy these BD tools to the cloud environment in an automated way. Thus, members of different research communities can contribute and share their tools or BD scripts within the BD framework using the TC. In the TC web user interface, users can provide citation information about their tool and will get proper credit for their effort in creating the software

A BD script that wraps the tool’s conversion capability and exposes it within BD service is known as converter; while a BD script that wraps the extraction capability of the tools and makes it available within the BD system is known as extractor. In subsequent subsections we will explain how to write an extractor or a converter through examples pertaining to archival data. In [7], creating of a BD tool (extractor or a converter) has been described in brief. To make writing of an extractor easy, a python package known as pyClowder was developed that handles common interactions with Clowder.

4.1 Create BD Tools

The CI-BER data observatory primarily consists of archival data, records from many federal agencies, cities, and civic organizations. These data are in many formats and in a variety of original folder structures. The unique challenge for the digital archivist at this scale is simply to know what they have in these collections and where, such that they can take appropriate preservation actions and provide access to researchers. We looked at the many extractors provided by the NCSA team, a compendium of computer vision and 3D modelling feature extractors, amongst others. We found that we needed additional extractors more germane to digital preservation practice, namely file characterization and digest. We created three extractors specific to archival data with the aim of applying them within CI-BER.

4.1.1 Siegfried extractor

The first extractor is based on Siegfried [8], which is a fast file characterization tool. It identifies sub-formats or format versions and other format characteristics about each file. These formats are discovered through byte matching with

[2] http://scape-project.eu/

[3] http://json-ld.org/

[4] opensource.ncsa.illinois.edu/bitbucket/projects/CATS/repos/pyclowder

[5] http://cassandra.apache.org/

[6] http://www.imagemagick.org/

[7] github.com/tesseract-ocr

Code 1: Connects to RabbitMQ with proper credentials

```
# connect to rabbitmq
extractors.connect_message_bus(extractorName =
     extractorName, messageType = messageType,
     processFileFunction = process_file,
     rabbitmqExchange = rabbitmqExchange,
     rabbitmqURL=rabbitmqURL)
```

the file patterns found in a registry of format signatures, PRONOM [9]. PRONOM, run by the National Archives of the United Kingdom, was the first web-based, public format registry in the world. Siegfried in particular is a project of Richard LeHane [8]. We use PRONOM-based file characterization in order to understand the exact format used in a file. The formats identified by file extension can be arbitrary, as data files can be renamed in arbitrary ways and as they often are unrecognized in older archival data. The Siegfried extractor is a BD tool that uses Siegfried, signature-based format identification tool, under the hood. Now provided as a BD extraction service, this is helping us obtain format data from the 100 million files that make up CI-BER.

To write a Siegfried extractor, we use the Clowder integration package for Python, pyClowder; and the Siegfried tool. Code snippet 1 shows the way to connect to RabbitMQ with proper credentials. Code snippet 2 shows the implementation of the *process_file* method based on the Siegfried tool and also how to upload the extracted metadata to the Clowder web app. The metadata extracted can be accessed using DTS API. Note the way Siegfried is called within the *process_file* method (Line 8). *connect_message_bus* and *upload_file_metadata_jsonld* are methods from pyClowder package.

4.1.2 FITS extractor

The second extractor is based on File Information Tool Set (FITS)[9], a file characterization toolkit. The FITS wraps several of the known digital preservation tools within it. They all run on a given file and the results are presented in one report where they can be compared, including points of agreement and disagreement. FITS is slower to run than Siegfried, but produces more data for analysis. It includes DROID, which does exactly the same PRONOM-based format identification as Siegfried. So a FITS file report will allow an archivist or an archival analytics tool to compare PRONOM identification with other tools, such as the Linux FileInfo tool, which has its own internal list of formats and byte patterns.

4.1.3 Byte Digest

The third extractor we created was for computing byte digests for files. We created a python based extractor that efficiently computes the MD5, SHA1, SHA256, SHA384, and SHA512 digests in a single pass through the data. Repositories rely on these kinds of digests to ensure the fixity of data across a variety of storage systems. Repositories may want to rely on the DTS for all forms of data extraction, or as a third-party cross-check to compare with digests created within proprietary systems. By comparing a locally computed digest with the digest coming back from the DTS, we can also ensure that the file data sent to the DTS was able to reach the extractors intact.

Code 2: BD script- Siegfried Extractor’s process file implementation and upload methods

```
1   # Process the file and upload the results
2   def process_file(parameters):
3       global extractorName
4
5       inputfile = parameters[’inputfile’]
6
7       # call the Siegfried (sf) program
8       resultStr = subprocess.check_output([’sf’,
             ’-json’, inputfile],
             stderr=subprocess.STDOUT)
9       result = json.loads(resultStr)
10
11      afile = result[’files’][0] # always one file
             only
12
13      content = {} # assertions about the file
14      content[’dcterms:extent’] = afile[’filesize’]
15
16      matches = []
17      for match in afile[’matches’]:
18          _logger.info(match)
19          m = {}
20          if ’id’ in match:
21              m[’@id’] = ’info:pronom/’+match[’id’]
22          if ’format’ in match:
23              m[’sf:name’] = match[’format’]
24          if ’version’ in match:
25              if len(match[’version’].strip()) > 0:
26                  m[’sf:version’] = match[’version’]
27          if ’mime’ in match:
28              m[’sf:mime’] = match[’mime’]
29          if ’basis’ in match:
30              m[’sf:basis’] = match[’basis’]
31          matches.append(m)
32
33      if len(matches) > 0:
34          content[’dcterms:conformsTo’] = matches
35
36      #wraps the metadata in JSON-LD format
37      jsonld_metadata = jsonld_wrap (content)
38
39      # upload metadata (metadata is a JSON-LD array
             of dict)
40      extractors.upload_file_metadata_jsonld(mdata =
             jsonld_metadata, parameters = parameters)
```

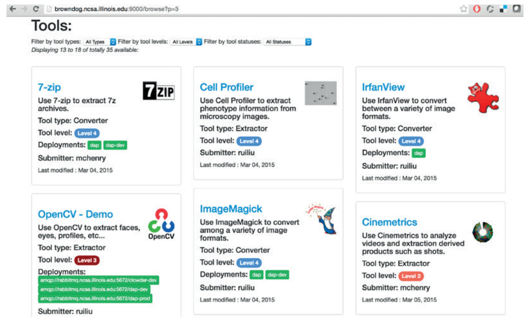
4.1.4 Imagemagick Converter

In this subsection we provide an example of a converter to be deployed within the BD system. We chose imagemagick, a third-party software with file format conversion capabilities. As described in [7] to write a converter, we provided in the comment of the script - line 2 : software name with version number, line 3: data type supported, i.e., image, in this case, line 4: list of input formats supported by imagemagick, line 5: list of supported output formats. Line 12 and 14 contain the actual imagemagick *convert* function call that converts an input file in supported input format to specific supported output format. For example, 3 script allows conversion of an image in pcd format to svg format.

[8] github.com/richardlehane/siegfried
[9] http://projects.iq.harvard.edu/fits

Code 3: BD Script - Imagemagick Converter

```
1   #!/bin/sh
2   #ImageMagick (v6.5.2)
3   #image
4   #bmp, dib, eps, fig, gif, ico, jpg, jpeg, jp2, pcd,
        pdf, pgm, pict, pix, png, pnm, ppm, ps, rgb,
        rgba, sgi, sun, svg, tga, tif, tiff, ttf, x,
        xbm, xcf, xpm, xwd, yuv
5   #bmp, dib, eps, gif, jpg, jpeg, jp2, pcd, pdf, pgm,
        pict, png, pnm, ppm, ps, rgb, rgba, sgi, sun,
        svg, tga, tif, tiff, ttf, x, xbm, xpm, xwd, yuv
6
7   output_filename=$(basename "$2")
8   output_format="${output_filename##*.}"
9
10  #Output PGM files as ASCII
11  if [ "$output_format" = "pgm" ]; then
12   convert "$1" -compress none "$2"
13  else
14   convert "$1" "$2"
15  fi
```



browndog.ncsa.illinois.edu:9000/browse?p=3

Tools:

Filter by tool types: All Types Filter by tool levels: All Levels Filter by tool statuses: All Statuses

Displaying 13 to 18 of totally 35 available:

7-zip
Use 7-zip to extract 7z archives.
Tool type: Converter
Tool level: Level 4
Deployments: dap dap-dev
Submitter: mchenry
Last modified : Mar 04, 2015

Cell Profiler
Use Cell Profiler to extract phenotype information from microscopy images.
Tool type: Extractor
Tool level: Level 4
Submitter: ruiliu
Last modified : Mar 04, 2015

IrfanView
Use IrfanView to convert between a variety of image formats.
Tool type: Converter
Tool level: Level 4
Deployments: dap
Submitter: ruiliu
Last modified : Mar 04, 2015

OpenCV - Demo
Use OpenCV to extract faces, eyes, profiles, etc...
Tool type: Extractor
Tool level: Level 3
Deployments:
amqp://rabbitmq.ncsa.illinois.edu:5672/clowder-dev
amqp://rabbitmq.ncsa.illinois.edu:5672/dap-dev
amqp://rabbitmq.ncsa.illinois.edu:5672/dap-prod
Submitter: ruiliu

ImageMagick
Use ImageMagick to convert among a variety of image formats.
Tool type: Converter
Tool level: Level 4
Deployments: dap dap-dev
Submitter: ruiliu
Last modified : Mar 04, 2015

Cinemetrics
Use Cinemetrics to analyze videos and extraction derived products such as shots.
Tool type: Extractor
Tool level: Level 2
Submitter: mchenry
Last modified : Mar 05, 2015

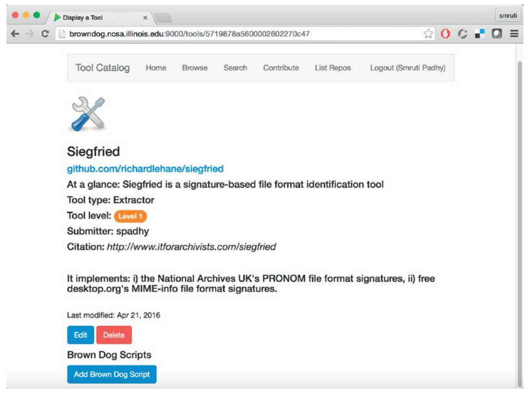
Figure 1: Tools Catalog web user interface showing list of tools/BD tool available in TC

4.2 Contribute and Share tool

To enable users to contribute and share a tool and its corresponding BD scripts, a Tools Catalog (TC) web application has been designed and is provided as a web service for BD users. Figure 1 shows the TC web user interface where a user can browse all tools information and BD tools/scripts that are being shared through TC, and can download BD scripts. It also has options to add tools information and contribute BD scripts and for admin to approve/disapprove a submitted BD tool/script. Figure 2 shows the specific tool information, e.g., Siegfried software information, after it has been added to TC.

5. INTEGRATED BROWN DOG TOOLS WITH REPOSITORIES

The DCIC team has integrated a number of services around the Indigo archival repository in Maryland. For demonstration purposes we have installed several Elasticsearch[10] nodes and the Kibana visualization tool[11]. In addition to rich search, these give us metrics and visualizations of the collections as they are enhanced with new data from Brown Dog.



Display a Tool

smruti

browndog.ncsa.illinois.edu:9000/tools/5719878a5600002602270c47

Tool Catalog Home Browse Search Contribute List Repos Logout (Smruti Padhy)

Siegfried
github.com/richardlehane/siegfried
At a glance: Siegfried is a signature-based file format identification tool
Tool type: Extractor
Tool level: Level 1
Submitter: spadhy
Citation: *http://www.itforarchivists.com/siegfried*

It implements: i) the National Archives UK's PRONOM file format signatures, ii) free desktop.org's MIME-info file format signatures.

Last modified: Apr 21, 2016

Edit Delete

Brown Dog Scripts

Add Brown Dog Script

Figure 2: Displays Siegfried tool information after it has been added to Tools Catalog using Add Tool form with proper citation.

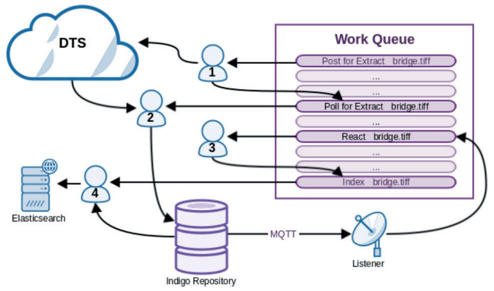


Figure 3: Workers and Task Queue

5.1 Simple Middleware

In order to coordinate the workflows we want around the Indigo repository, Brown Dog services, and Elasticsearch, we needed to create additional middleware. This middleware is mostly a work queue system, which allows us to perform work asynchronously, and at a predictable rate, instead of having to perform all operations immediately upon deposit or immediately in response to Indigo data changes.

When a user makes a change to the data in the Indigo repository, say they upload a new file, a message is generated and broadcast to any software that is listening to the Indigo message feed. The first step in the DCIC workflow is to listen for these Indigo messages. Our listener converts each message into a task and places that task in our work queue. The first task, called React, can be described as “respond to this Indigo message”. Our simple listener has no trouble keeping up with all of the Indigo changes, as the significant work has been postponed for later.

Next in the workflow we have a pool of workers, see Figure 3. These are software processes that are waiting to pick up and perform any work from the work queue above. At the DCIC we normally have ten workers running, but if the queue of work keeps growing due to a large number of deposits, we can increase the number of workers. Workers can be distributed across multiple servers if necessary. There may be a variety of tasks added to the work queue and these workers can perform any of them. Let’s look at some of the tasks that make up our workflow.

[10] www.elastic.co
[11] www.elastic.co/products/kibana

5.2 Tasks That Support Workflow

- **React** - Responds to the content of an Indigo message. This task is created by the Indigo listener. The worker will figure out what change was made in Indigo and what tasks should be performed as a result. For instance, any metadata change will result in an index task. This task adds other tasks to the work queue.
- **Index** - Indexes (or re-indexes) a file or folder in Elasticsearch. The worker will fetch any necessary data from Indigo.
- **Deindex** - Removes a file or folder from Elasticsearch.
- **Post for Extracts** - Uploads a file to Brown Dog's Data Tilling Service (DTS) for feature extraction. Adds a Poll for Extracts task to the queue.
- **Poll for Extracts** - Checks to see if the DTS extract operations above are complete yet. If incomplete, this task is scheduled to run again after a delay. If complete, the worker downloads the extracted metadata from DTS and puts this metadata into Indigo.
- **Text Conversion** - Uploads a file to the DAP for conversion into a text file, if possible. Schedules a Poll for Text task to run after a delay.
- **Poll for Text** - Checks to see if text conversion is available from DAP yet. If text is not yet available, this task is scheduled to run again after a delay. If text is available, it downloads the text and puts the text into a full text field in Indigo metadata.

Each step in the workflow is separated into a discrete task and each task is only performed when it gets to the front of the work queue. We can monitor the work queue to see how long it has become and how long tasks must wait to be performed. Organizations with available server resources can scale up the number of workers to keep up with demand. Organizations with few server resources can control server load and still eventually process the queued jobs.

The asynchronous middleware we describe here was implemented in the Python language and uses Celery[12] task queues. The work queues that are managed by Celery are persisted in a RabbitMQ messaging service. Each task may be relatively simple. For example Code 4 shows the complete code for the React task.

As shown, the React task schedules other tasks on repository paths in response to the Indigo operation. Other tasks are longer and involve requests to web services, either Brown Dog, Indigo or Elasticsearch. In some cases further workflow steps will result indirectly, via calls to Indigo services. For example, after full text is added to an Indigo metadata field, then the listener will be notified and will schedule a React task, then the React task will schedule an Index task, which will update Elasticsearch to include a full text field.

5.3 Workflow On Demand

As we incorporate more services into the workflow, or add new fields to our Elasticsearch index, we will add new workflow reactions in the React task. These will respond to new file deposits and changes in the data. However, we also want to trigger these new workflows on existing repository data. For this we turn to a special task called Traverse:

Code 4: Code Sample for the React Task

```
@app.task
def react(operation, object_type, path,
     stateChange):
 if 'create' == operation:
      index.apply_async((path,))
      if 'resource' == object_type:
          postForExtract.apply_async((path,))
 elif "update_object" == operation:
      index.apply_async((path,))
 elif "delete" == operation:
      deindex.apply_async((path, object_type))
```

- **Traverse** - Traverse schedules another task for each file or folder within a given part of the repository tree, starting at the root and extending in a breadth-first manner to the branches. Traverse lets you perform workflow on demand for large areas of the repository. Traverse works recursively, making use of the work queue to schedule a further Traverse task for each subfolder at a given level. Recursive traverse tasks can reliably process folders of arbitrary depth without any long running worker processes.

For instance, if we add new fields to our Elasticsearch, we will traverse the entire repository to apply the Index task. If we add a new workflow, such as conversion to plain text, to the React tasks, we can apply the new workflow to existing data through the Traverse task. We add Traverse tasks to the queue directly, via a command-line script, rather than through the Indigo listener.

5.4 Taking Incremental Steps

With the 100 Million files in CI-BER collections we approach the problems of billion-file scale. Even given an asynchronous work queue, if we traverse a large collection without pausing, we will quickly overload the work queue with pending tasks, bringing the machine it runs on to a halt. Instead a traverse must proceed in stages. The traverse task has special logic that checks the length of the pending task queue and postpones itself whenever the queue is too large. A traverse will only proceed with creating more tasks while the queue is of a manageable size. In Figure 4 you can see the size of the overall queue over time, including all tasks, as we gradually traversed a collection. Each bump was created by a traverse operation that waited until the queue was small and then added more tasks.

In this way the workers can gradually traverse the entire repository to bring all materials up to date with respect to the current workflow.

6. VISUALIZATION OF EXTRACTED METADATA

The integration between Indigo, Brown Dog, and Elasticsearch creates an expanded set of metadata fields in the repository. When these are indexed in Elasticsearch, we can ask new questions and understand the collections and folders at every level in greater level of detail. All of the following charts were created in the Kibana visualization tools for Elasticsearch, using data drawn from Brown Dog services. Each chart is part of the overview provided by a Kibana dashboard. The Kibana dashboard, our overview, can be redrawn with arbitrary index fields as filters, much like the drill-down feature of a faceted search system. Most importantly we can look at the dashboard of visualizations for any folder in CI-BER to better understand the contents.

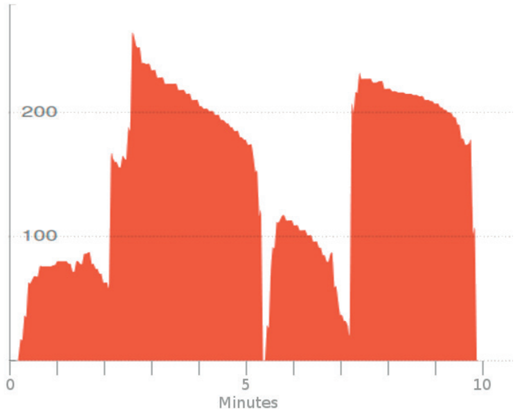
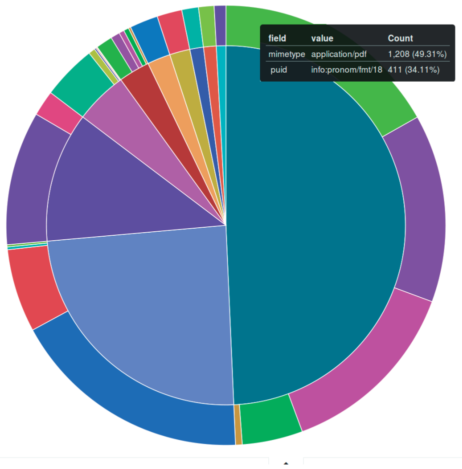


Figure 4: Incrementally Adding Work to the Task Queue over Time



field	value	Count
mimetype	application/pdf	1,208 (49.31%)
puid	info:pronom/fmt/18	411 (34.11%)

Figure 5: Kibana visualization of mimetype (inner sections) and format (outer sections)

6.1 Format Distribution

One of the insights we gain from the Siegfried extractor is detailed format information for every file we process. This chart captures a high level view of the most common file formats in the repository. The concentric pie chart shows mimetypes in the inner circle and then breaks these mimetypes down into specific sub-formats in the outer ring.

In its web-based interactive form, this chart has pop-up labels and can be used to target further preservation and access enhancements, such as format migration or format specific extraction and indexing. In the collection shown above the most common mimetype is application/pdf, with a distribution of subformats from PDF v1.2 through v1.6.

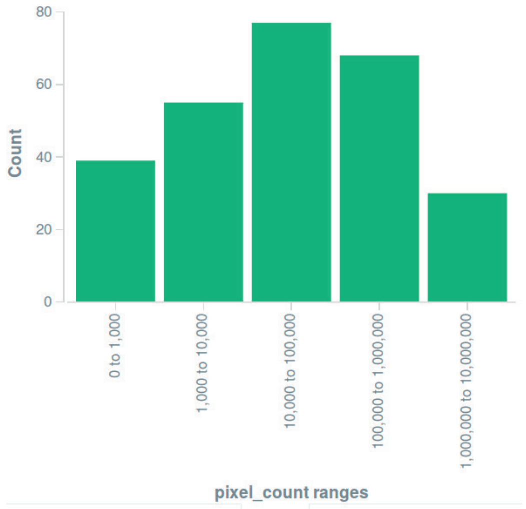


Figure 6: Total Images by Pixel Count (Orders of 10)

6.2 Image Features

Our Elasticsearch cluster includes fields that the DTS has derived from images. Using these metrics we can formulate queries based on image content. For instance we can easily formulate an Elasticsearch query that will find all megapixel images. Below we have a graph showing the numbers of images in a collection falling within pixel count ranges, in orders of ten.

There are a number of other visual features extracted by Brown Dog that may be useful. Visual symmetry is reflected in skewness and kurtosis factors. Human features are tagged and delimited by box regions, including faces, eyes, profiles, and close-ups. Note that this will include both photographs and realistic drawings of people. By indexing these features, we can find all of the images that feature people, or that feature a certain number of people.

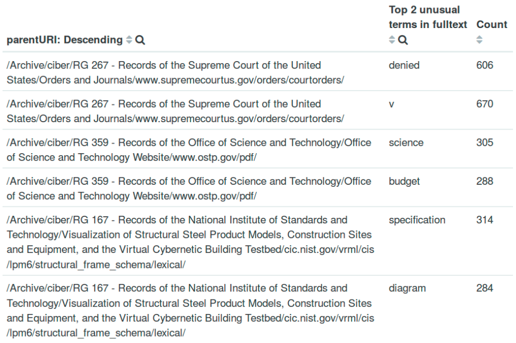
6.3 Textual Features

We leveraged both the OCR and format conversion offerings of Brown Dog to acquire as much of the text content from the CI-BER files as possible. The text was recorded in Indigo metadata fields and indexed by Elasticsearch. We have done little beyond a search index with these text fields so far, but we see much more potential for text analysis, now that the text is no longer locked in a file format. For one example, we can find unusual terms to understand how text in one part of our repository is different from elsewhere.

In the table generated by Kibana above there are two rows for every folder, we see that the most unusual terms in U.S. Supreme Court documents are, unsurprisingly, "denied" and "v", as in "motion was denied" and "Marbury v. Madison". The OSTP is more concerned with "science" and "budget", while NIST is more concerned with "specification" and "diagram". Unusual terms is an especially interesting approach for archival material because it is comparative, showing those terms that are distinctive for each "bucket" within the result set. You can define what your "buckets" are through any other indexed field, be it the folder path, the author, or the year files were created.

The Kibana portal used to create the graphs above exists

[12] http://www.celeryproject.org/



parentURI: Descending ⇕ 🔍	Top 2 unusual terms in fulltext ⇕ 🔍	Count ⇕
/Archive/ciber/RG 267 - Records of the Supreme Court of the United States/Orders and Journals/www.supremecourtus.gov/orders/courtorders/	denied	606
/Archive/ciber/RG 267 - Records of the Supreme Court of the United States/Orders and Journals/www.supremecourtus.gov/orders/courtorders/	v	670
/Archive/ciber/RG 359 - Records of the Office of Science and Technology/Office of Science and Technology Website/www.ostp.gov/pdf/	science	305
/Archive/ciber/RG 359 - Records of the Office of Science and Technology/Office of Science and Technology Website/www.ostp.gov/pdf/	budget	288
/Archive/ciber/RG 167 - Records of the National Institute of Standards and Technology/Visualization of Structural Steel Product Models, Construction Sites and Equipment, and the Virtual Cybernetic Building Testbed/cic.nist.gov/vrml/cis /lpm6/structural_frame_schema/lexical/	specification	314
/Archive/ciber/RG 167 - Records of the National Institute of Standards and Technology/Visualization of Structural Steel Product Models, Construction Sites and Equipment, and the Virtual Cybernetic Building Testbed/cic.nist.gov/vrml/cis /lpm6/structural_frame_schema/lexical/	diagram	284

Figure 7: Unusual Terms in Folders Containing the Most Texts

as a separate analytics application that is not integrated into the access repository. When we bring these additional search and analytics features into our access repository, they will provide an overview in the context of the collection structure. We will render a dashboard on demand for any folder or collection in the repository, showing analysis of the contents. This brings a pre-made set of relevant analytics into view for every repository user, not just the person crafting the visualizations.

7. DISCUSSION

The workflows described here and their application to appraisal and access are in the early stages. There are several direct steps we will take next to further explore our study.

We have begun to scratch the surface of the data in CI-BER, running the extractors, etc. on a sampling of our collections. However, we have not run the workflow on the bulk of the scientific data in CI-BER, which will pose different challenges and opportunities. The Elasticsearch index and Kibana visualization tool, give us significant analysis features "out of the box" and have promise as an investigative tool for born digital materials, but the dashboards are not integrated into our user-facing access interface. Finally, we can connect repository users to the DTS Clowder item and collection interface, which delivers the complete superset of extracted data for each file, unfiltered by our local repository design and indexing choices. With these straightforward next steps we will improve our understanding of the potential for Brown Dog.

Another avenue to explore is the looping of data through DTS and DAP to extract more knowledge. For instance, we can first convert a document into full text via DAP, then feed the full text into the DTS for all manner of text analysis extractions, including natural language processing to discover dates and the names of people and places. The same text analysis can be applied to OCR text or transcripts extracted from audio. This text mining across diverse formats is hard to achieve traditionally, requiring a dedicated repository and software effort. Within the Brown Dog framework we may be able to bring it within reach of more institutions. A similar combining of Brown Dog services can be used to split out and process sections of files, such as the detailed content items within an MBOX, ZIP or disk image file.

The DTS provides us with metadata in the form of JSON-LD graphs. Presently we only pull certain field values from the JSON-LD, treating it as JSON. A triple store or graph database can be used to index all of the extracted data, from all of the files, in a larger graph. A graph of all of the extracted data opens the door to graph reasoning across the collections. For instance, you might establish that a set of people were working in a team for a time, since they have frequently corresponded or shared authorship on documents. Furthermore, a linked data store allows you to coordinate and query your local data alongside linked data in other places, such as dbpedia[13] One simple example is to link recognized place names with their matching resource in Geo Names. This gives you the ability to query for and index all files that pertain to any level of administrative region on a map. For example a document that mentions "Brooklyn" could be discovered via New York City and New York State.

8. CONCLUSION

We have demonstrated a model architecture, consisting of cloud-based Brown Dog services, Maryland/DCIC middleware, the Indigo repository, and the Elasticsearch applications, that function together at scale to populate the CI-BER collections with enriched metadata records.

We contributed our own extractors to Brown Dog, adding key digital preservation functions. We deployed the Siegfried extractor into the DTS, wrapping the functions of the Siegfried format identification software. While contributing the extractor required programmer effort, the integration of the extracted data into workflows was automatic, as Siegfried's format-related findings merged with the rest of our DTS-supplied metadata. The only change to the Maryland workflow was to decide which Siegfried data to put in the search index. This experience further shows us that Brown Dog is a potent aggregator of extraction and migration tools under one API, capable of multiplying the value of the tool building efforts in the broad data curation community.

Lastly, we find that an enriched supply of metadata directly extracted from digital materials can yield tremendous benefits in the analysis of collections. Data analytics software, such as Kibana, can be used without much domain-specific configuration to gain insight into collection contents. This gives us our first glimpse of what we can do with the expanding workflows and metadata.

Acknowledgments

This research & development has been funded through National Science Foundation Cooperative Agreement ACI-1261582. We would also like to thank the following partners and software projects: Archive Analytics Solutions, Inc. - creators of the Indigo Repository software; University of Maryland Division of IT - system administrator support; Richard LeHane - creator of the Siegfried Software.

References

[1] NCSA Brown Dog. http://browndog.ncsa.illinois.edu/, 2013 (accessed April 21, 2016).

[2] J. R. Heard and R. J. Marciano. A system for scalable visualization of geographic archival records. In *Large Data Analysis and Visualization (LDAV), 2011 IEEE Symposium on*, pages 121–122, Oct 2011.

[3] L. Marini, R. Kooper, J. Futrelle, J. Plutchak, A. Craig, T. McLaren, and J. Myers. Medici: A scalable multimedia environment for research. *The Microsoft e-Science Workshop*, 2010.

[4] K. McHenry, R. Kooper, and P. Bajcsy. Towards a universal, quantifiable, and scalable file format converter. *The IEEE Conference on e-Science*, 2009.

[5] K. McHenry, R. Kooper, M. Ondrejcek, L. Marini, and P. Bajcsy. A mosaic of software. *The IEEE International Conference on eScience*, 2011.

[6] J. Myers, M. Hedstrom, D. Akmon, S. Payette, B. Plale, I. Kouper, S. McCaulay, R. McDonald, A. Varadharaju, P. Kumar, M. Elag, J. Lee, R. Kooper, and L. Marini. Towards sustainable curation and preservation: The sead project's data services approach. *Interoperable Infrastructures for Interdisciplinary Big Data Sciences Workshop, IEEE eScience*, 2015.

[7] S. Padhy, G. Jansen, and etal. Brown dog: Leveraging everything towards autocuration. In *IEEE Big Data*, 2015.

[8] Richard Lehane. Siegfried: A signature-based file format identification tool. http://www.itforarchivists.com/siegfried, 2014 (accessed April 21, 2016).

[9] The UK National Archives digital preservation department. PRONOM. http://www.nationalarchives.gov.uk/PRONOM/, (accessed April 21, 2016).

[13] http://linkeddata.org/

Navigating through 200 Years of Historical Newspapers[*]

Yannick Rochat
Digital Humanities Laboratory
CH-1015 Lausanne
yannick.rochat@epfl.ch

Maud Ehrmann
Digital Humanities Laboratory
CH-1015 Lausanne
maud.ehrmann@epfl.ch

Vincent Buntinx
Digital Humanities Laboratory
CH-1015 Lausanne
vincent.buntinx@epfl.ch

Cyril Bornet
Digital Humanities Laboratory
CH-1015 Lausanne
cyril.bornet@epfl.ch

Frédéric Kaplan
Digital Humanities Laboratory
CH-1015 Lausanne
frederic.kaplan@epfl.ch

ABSTRACT

This paper describes the processes which led to the creation of an innovative interface to access a digital archive composed of two Swiss newspapers, namely *Gazette de Lausanne* (1798–1998) and *Journal de Genève* (1826–1998). Based on several textual processing steps, including lexical indexation, n-grams computation and named entity recognition, a general purpose web-based application was designed and implemented ; it allows a large variety of users (e.g. historians, journalists, linguists and the general public) to explore different facets of about 4 million press articles spanning an almost 200 hundred years period.

Keywords

Digital humanities, historical newspapers, innovative interface, language evolution, named entity recognition

1. INTRODUCTION

Newspapers are essential sources for the exploration of the past [4]. From a historical point of view, they document aspects and events of our societies from the perspective of contemporary actors and, from a linguistic point of view, they constitute (once digitized) large corpora that can be used to e.g. investigate the evolution of language(s). Both researchers and the general public benefit from online access to cultural heritages such as newspaper archives [15].

Many newspapers digitisation projects[1] have been realised in the last ten years [16, 22] thanks to the facilitated acquisition of larger storage amenities and higher computing power. Most projects provide access to the scanned documents but do not offer more than basic search through the textual content.

In Switzerland, the Swiss National Library has contributed to the digitisation of more than thirty newspapers.[2] The library centralises some of these projects[3], while others are hosted by public or private partners.[4]

In 2008, all original issues of the three journals composing the archives of *Le Temps*[5]–*Gazette de Lausanne*, *Journal de Genève* and *Le Nouveau Quotidien*[6] (1991–1998)–were digitised and made available for consultation to the public through a website.[7] Texts have been extracted from scanned pages using optical character recognition (OCR) and layout detection algorithms, allowing visitors to search through a corpus composed of close to 1 million pages and 4 million articles[8], covering 200 years of local, national and global news as seen from the French part of Switzerland.[9]

This article describes a web application offering a new interface to navigate this 200 year corpus. It was developed as part of a collaboration between *Le Temps*, the Swiss National Library and the Digital Humanities Laboratory of the Swiss Federal Institute of Technology of Lausanne (EPFL). Features formerly available like lexical search, editable time intervals or the possibility to look for a given issue based on the date were implemented. An image viewer that situates articles in their original contexts was developed allowing to browse full newspaper issues from the first to the last page without leaving the interface. Each page can be zoomed into up to a level allowing to see small details of graphics or comfortable on-screen reading. In addition, two methods stemming from natural language research to improve the navigation in the corpus, namely n-grams viewing and named entities, were adapted.

The remainder of the paper is organised as follows. In section 2 we describe the corpus composed of the two main newspapers, *Gazette de Lausanne* and *Journal de Genève*. Next, we present the text processes applied on the digital archive with the computation of n-grams (section 3) and the recognition of named entities (section 4). In section 5 we detail theoretical and technical aspects of the public interface and finally we conclude and consider future work in section 6.

2. LE TEMPS CORPORA

In this section, we present a few quantitative descriptors for this corpora (publication frequency, statistics of words and pages), then we display front pages for key moments in the history of these newspapers and sketch their stylistic evolution over time. Eventually, we discuss the encoding of the data.

2.1 General Statistics

Gazette de Lausanne and *Journal de Genève* reached regular and similar publication frequencies in the 1850s. Before that time, the situation was less harmonious. *Gazette de Lausanne* appeared rather regularly, around 100 times a year from 1804 to 1846 (see figure 1) while the number of issues per year of *Journal de Genève* varied from 52 issues (1828) to 246 issues (1834) between 1826 and 1850 (see figure 2).

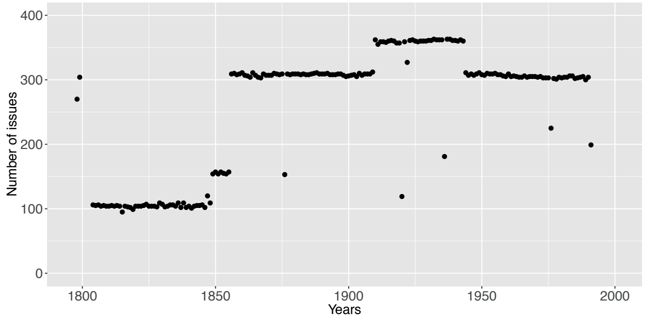


Figure 1: The number of issues per year of *Gazette de Lausanne*.

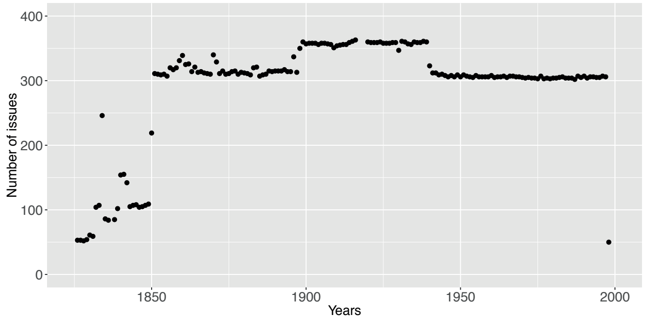


Figure 2: The number of issues per year of *Journal de Genève*.

There are eight outlying years in our dataset for *Gazette de Lausanne* (described in table 1, and no equivalent for *Journal de Genève*. With the exception of years 1798 and 1799, which are composed of issues from *Gazette de Lausanne*'s ancestors, it appears that these outliers are mostly years with missing data inherited from the original data set.[10] The task of retrieving the parts currently lacking is ongoing.

In our corpus, there are in total 441′579 printed pages for *Gazette de Lausanne* from 1798 to 1998[11] and 495′986 for *Journal de Genève* from 1826 to 1998. In addition, figures 3 and 4 show the average number of pages per issue for these two newspapers. With exception of the very first years and 1830s for *Gazette de Lausanne*, both newspapers were printed on 4 pages (one large sheet of paper, folded) until 1900s for *Journal de Genève* and 1940s for *Gazette de Lausanne*. Then the number climbed with a slowing down in the 1970s for both newspapers.

Table 1: Outlying years from figure 1.

Year	# of published issues
1798	270
1799	304
1876	153
1920	119
1922	320
1936	181
1976	225
1991	199

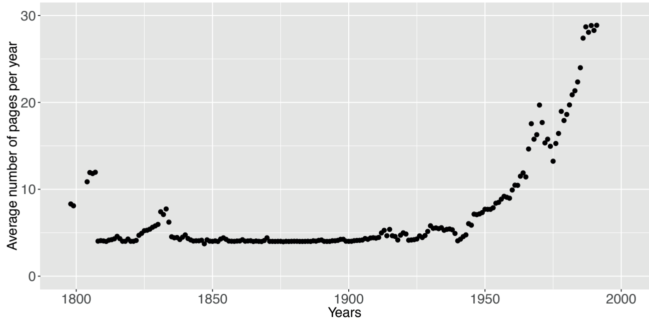


Figure 3: Average number of pages by issue, per year, in *Gazette de Lausanne*.

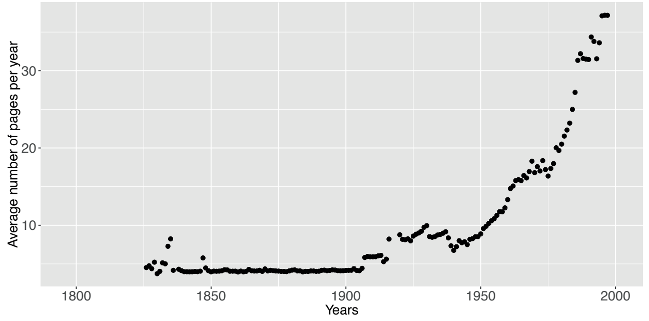


Figure 4: Average number of pages by issue, per year, in *Journal de Genève*.

Selected front pages from *Gazette de Lausanne* and *Journal de Genève* at births (see figures 12, 13 and 14) and deaths of these journals (see figures 15 and 16) are shown in the appendix.

[*] Supported by the Swiss National Library.

[1] http://en.wikipedia.org/wiki/Wikipedia:List_of_online_newspaper_archives Accessed on April 24th, 2016.

[2] http://www.nb.admin.ch/themen/02074/02076/03887/?lang=en Accessed on April 24th, 2016.

[3] http://newspaper.archives.rero.ch/ Accessed on April 24th, 2016.

[4] http://www.nb.admin.ch/public/04506/04514/index.html?lang=en Accessed on April 24th, 2016.

[5] A Swiss newspaper launched in 1998.

[6] At the time of writing, the inclusion of *Le Nouveau Quotidien* in the new website is ongoing.

[7] It will be removed in the future. At the time of writing, it is accessible at old.letempsarchives.ch

[8] Including images with captions, and advertisements.

[9] These newspapers were written in French.

[10] For example, all issues are missing : from 1800 to 1803, from July 1876 to December 1876, from May 1920 to December 1920, from January to June 1936.

[11] During years 1991 to 1998, the two newspapers were merged into a single one whose name was *Journal de Genève et Gazette de Lausanne* (see figure 16).

2.2 Encoding

The whole archive, including text and images, weighs 22 terabytes. The structure of each newspaper issue is encoded in a master XML file which lists articles[12] along with metadata information (e.g. article boxes and ordering, used font, etc.). In turn, the content of each article, that is to say words and their positions, is encoded in a proper XML file. Besides the XML text material, each page image is saved in TIFF format and all pages of an issue are saved in a single PDF containing the OCRed text.

The quality of the OCR has not been evaluated at this stage, but some mistakes are immediately noticeable, mostly due to bad conservation of paper, to the digitisation process (transparency, creases, stains), and to ink drips at the time of printing, all common phenomena in this type of project.

3. N-GRAMS

An *n-gram* is an ordered sequence of n consecutive words. For instance, given the phrase *La Gazette de Lausanne*, *La Gazette*, *Gazette de* and *de Lausanne* are 2-grams, whereas each word taken separately is a 1-gram.

N-grams were extracted from the XML text files, by considering alphanumeric tokens only. In order to compute the n-grams relative frequencies, we chose a time granularity of one year and divided the number of occurrences of each n-gram by the total number of n-grams occurrences for the same period. Obviously, as n increases, so does the likeliness of any n-gram to be unique, and thus the number of distinct n-grams converges towards the number of words in the corpus. For that reason, storing the n-grams becomes rapidly costly in terms of volume for large values of n.

Visualising n-grams frequency distributions on a given corpus allows to test hypotheses about linguistic and sociolinguistic evolutions, as preceding works demonstrated [18, 23]. In order to help users gather knowledge for a given query on the whole corpora, a viewer allowing to display the variations of n-grams relative frequencies over time was created. Examples of this n-gram viewer can be seen in figures 5 and 6. N-grams distributions are influenced by linguistic and socio-cultural factors, but also by constraints related to the journals themselves (e.g. the diversity of covered topics, article sizes, etc.). As an example, the behaviour of the n-gram *1914* is greatly impacted by the First World War, which is not a linguistic factor. On the other hand, the lexical diversity might be influenced by the length of articles as well as linguistic evolution. All these factors contribute, in different proportions, to the n-grams frequencies evolution and have to be considered together.

Uses of n-grams are manifold. From a set of n-grams distributed over time, we can extract linguistic, semantic and sociocultural information. Several researches are currently underway and use the extraction of absolute and relative frequencies of n-grams. For example, a study explored the different typologies of n-grams curves identifying core processes and classifying n-grams in these archetypical categories [6]. This study considered the question of reversing the n-gram viewer paradigm, searching in the space of n-grams frequencies curves instead of searching in the space of n-grams. Another study defined the notion of n-gram cores and resilience, allowing to compare corpora and study linguistic evolution through the concept of words resilience instead of linguistic changes [5].

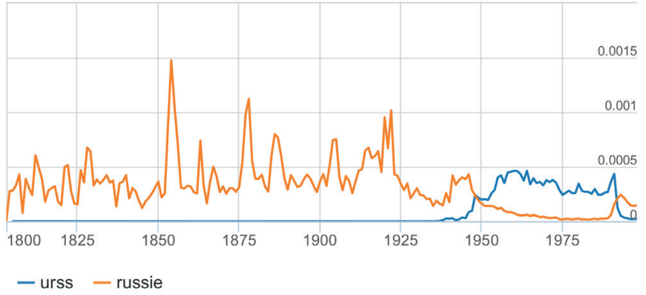


Figure 5: Visualisation of 1-grams “russie” (Russia) and “urss” (USSR).

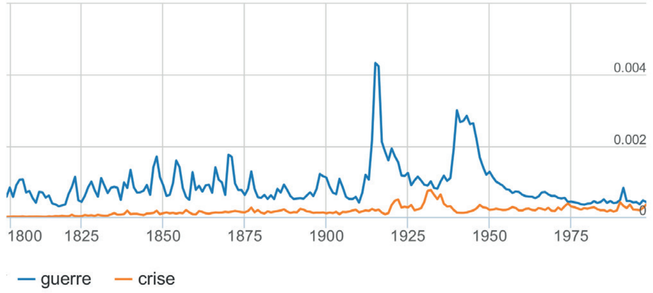


Figure 6: Visualisation of 1-grams “guerre” (war) and “crise” (crisis).

4. NAMED ENTITIES

Recognition and processing of real-world entities is essential for enabling effective text mining. Indeed, referential units such as names of persons, organisations and locations underlie the semantics of texts and guide their interpretation. Known as named entities (NE), these units are major bearers of information and can help answering the questions of *Who did What to Whom, Where and When*? (known as the *5Ws* in journalism). First introduced during the 6th Message Understanding Conference [12], named entity processing have evolved significantly over the last two decades, from entity recognition and classification to entity disambiguation and linking[13] [10, 21]. More recently, NE processing has been called upon to contribute to the research area of Digital Humanities where algorithms have to deal with OCRed documents [25, 26] and languages and documents of earlier stages [7, 11, 30].

In the context of designing and developing a new interface to enable users to search through two of the newspapers composing *Le Temps* archive, implementing a named entity recognition system appeared as an obvious desideratum. Although many NE processing tools are now available almost “off-the-shelf”, they can hardly be applied on *Le Temps* documents for various reasons. Tools developed by private companies (e.g. Open Calais[14], Zemanta[15], Alchemy[16]) are most of the time intended for English language and, when available for French, are only accessible through limited web services–a framework unsuitable when analysing millions of documents. Moreover, APIs and tag sets (i.e. named entity categories) of those tools are regularly updated, which results in undesirable maintenance problems. On the academic side, various entity linking tools are being developed by the Natural Language Processing and Semantic Web communities. DBpedia Spotlight [8, 17], AIDA [32] and BabelFy [20] are dedicated to the spotting of entity mentions in texts and their linking to entities stored in knowledge bases (KBs). If they are able to assign referents to entities in text (i.e. entity disambiguation), these tools do not however perform real named entity recognition in the sense that they can only spot names of entities which are present in the given KB. Besides, background KBs are for the most part derived from Wikipedia and thereby contain primarily VIPs, which is unsuitable for recognising the *John Doe(s)* of past and present days from *Le Temps* collection. Finally, those tools are well developed and maintained for English language; it is possible to deploy them on new languages but it requires a huge effort for a result which might not meet all needs.

Without discarding the option of using one of these tools at a later stage, as of now we sought a solution able to (1) parse French language, (2) recognise all entity mentions, and (3) be executed offline. To this end, we used a rule-based system using the ExPRESS formalism [24] such as deployed by the *Europe Media Monitor* (EMM) [29] for multilingual NER [28]. ExPRESS is an extraction pattern engine based on finite state automata. It allows to define rules or patterns which, coupled with appropriate lexical resources and preprocessing (tokenization and sentence splitting), can detect and type specific phrases in texts. Named entity recognition is implemented via a cascade of grammar files where units are detected and processed in increasing order of complexity. In concrete terms, NE rules focus on typical patterns of person, location and organisation names, e.g. an adjective (*former*) followed by a function name (*President of the Confederation*), a first (*Ruth*) and a last (*Dreifuss*) name. Units such as *former* and *President* are called trigger words; besides modifiers and function names they cover professions (*guitarist, football player*), demonyms and markers of religion or ethnical groups (*Italian, Genevan, Bambara, Muslim*), expressions indicating age (*42 years-old*), and more. It is worth noticing that this system performs named entity recognition and classification but not disambiguation.

We applied our named entity grammars[17] on articles of *Le Temps* archive for the recognition of Person and Location names (we reserve the Organisation type for future work). In order to speed up the process and ease the debugging, we executed our process in parallel on a very powerful computing node (48-core, 256GB of RAM). Parsing of all files took a couple of hours. In order to allow maximum flexibility with the usage of data, processing results are first stored in JSON[18] format. They are afterwards converted in the Resource Description Framework (RDF) so as to allow final data publication as Linked Data [2, 13]. The ontology used to represent extracted entities revolves around two core elements, *Article* and *EntityMention*, each one being further qualified with specific properties. We made use of classes and properties defined by the Dublin Core terms[19], NIF[20], OLiA[21] and LexInfo[22] vocabularies. The RDF graph is loaded on a triple store (Virtuoso open source) whose SPARQL endpoint is available from the interface, as we shall see in the next section. Users can access about 30 million entity mentions of type Location and 20 million of type Person. Thanks to the extraction of detailed information along with person names and to their RDF representation, it is possible to explore various dimensions of person entities. Examples of queries against the data set include :

- all person mentions having a specific function (e.g. *German Chancellor*) in articles issued between date x and date y;
- all functions of a specific person mention ordered chronologically, with the possibility to get the source articles;
- all articles mentioning conjointly 2 or more specific person mentions;
- all person mentions which occur with a specific title or function;
- etc.

Future developments regarding this text processing module involve NER evaluation, processing of Organisation entities and entity disambiguation.

5. WEB APPLICATION

5.1 Interface Principles

The interface design we addressed typically falls under a lack of known or typical use cases. With any website, we expect the base of users and their expectations to be very wide and diverse. Regarding the old website, the only statistic we could use would have been the user search history. However, this tells little about their intents and we ignore if the information they found was relevant to them.

We thus needed to define a set of basic requirements that would follow the most generic possible use cases, yet providing modern and powerful features to journalists, historians and information scientists. The core features that were outlined by preceding studies on similar archives are the following :

- A global, full-text, high performance search engine is generally the preferred way to access information, both to novice and expert users [9]. The added value of finding aids such as advanced search options or a hierarchical organisation is however subject to debate [31].
- Articles should always be read in their full publication context [4].
- Each page needs to be easily referenced by an unique URL, so it can be quickly stored for later access in a situation of information gathering [1].

In addition to the search engine, we needed to come up with an appealing way to browse search results. Unfortunately, no relevance score can easily be derived from the way

[12] The word “article” represents articles, images and advertisements.

[13] Entity linking corresponds to the task of linking an entity metnion to a unique identifier in a knowledge base, e.g. DBpedia.

[14] www.opencalais.com

[15] www.zemanta.com

[16] www.alchemyapi.com

[17] Composed of ca. 130 rules for Person and Location names.

[18] JavaScript Object Notation.

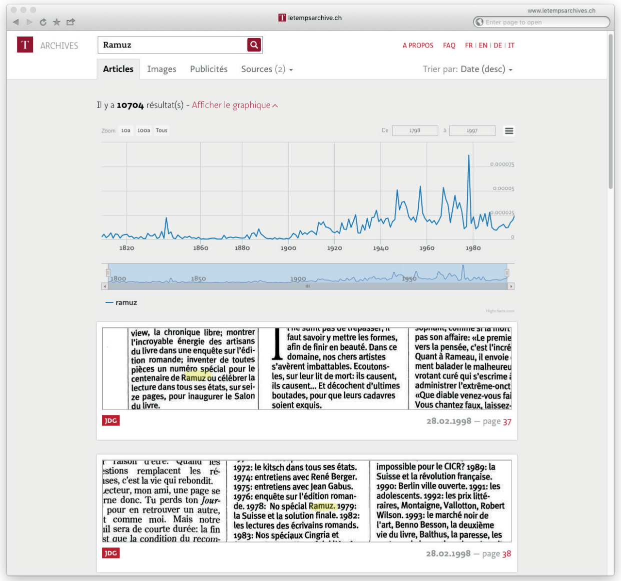
[19] http://purl.org/dc/terms/

[20] http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#

[21] http://purl.org/olia/olia.owl#

[22] http://www.lexinfo.net/ontology/2.0/lexinfo#

contents are organised, as there are no links between articles that may allow to guess their relative importance, nor a clear way to predict which kind of content might be interesting. To answer this question, we thus introduced the n-gram visualisation as a very part of the search results. In this manner, results of search queries consist of, first, the n-gram viewer featuring the evolution of query term ies over time on both journals and, second, snippet previews of retrieved articles. The n-gram viewer allows users to get a quick hint at periods of interest and to select more precise time frames to dig for interesting results. Figure 7 shows how typical search results are presented.



Il y a **10704** résultat(s) - Afficher le graphique

view, la chronique libre; montrer l'incroyable énergie des artisans du livre dans une enquête sur l'édition romande; inventer de toutes pièces un numéro spécial pour le centenaire de Ramuz ou célébrer la lecture dans tous ses états, sur seize pages, pour inaugurer le Salon du livre.

faut savoir y mettre les formes, afin de finir en beauté. Dans ce domaine, nos chers artistes s'avèrent imbattables. Ecoutons-les, sur leur lit de mort: ils causent, ils causent… Et décochent d'ultimes boutades, pour que leurs cadavres soient exquis.

28.02.1998 — page 37

1972: le kitsch dans tous ses états. 1974: entretiens avec René Berger. 1975: entretiens avec Jean Gabus. 1976: enquête sur l'édition romande. 1978: No spécial Ramuz. 1979: la Suisse et la solution finale. 1982: les lectures des écrivains romands. 1983: Nos spéciaux Cingria et

impossible pour le CICR? 1989: la Suisse et la révolution française. 1990: Berlin ville ouverte. 1991: les adolescents. 1992: les prix littéraires, Montaigne, Vallotton, Robert Wilson. 1993: le marché noir de l'art, Benno Besson, la deuxième vie du livre, Balthus, la paresse, les

28.02.1998 — page 38

Figure 7: Search results. At the top, period selection using the n-gram viewer. At the bottom, previews from the found articles.

The necessity of viewing full pages with an adequate resolution called for a tailored solution. The technical requirements are as follows :

— The search engine results need to access previews that can be anywhere in the pages, typically showing the found word(s) in a context of a couple sentences.

— The high quality scans of the full pages are too big to be loaded as they are[23], yet we need to be able to present them as a whole to the user and to enlarge the relevant parts, possibly up to the highest definition available.

— All the images sent out to the client must be optimised to keep low loading times and acceptable server loads.

Those can be addressed in a nice way using a web image server supporting multiple image formats (raw files for archiving and preservation and optimised ones for web delivery) and tiling. We selected an image server responding to the International Image Interoperability Framework (*IIIF*) norm, for its outstanding interoperability and academic approach [27].

Figures 8 and 9 illustrate the use of the viewing interface.

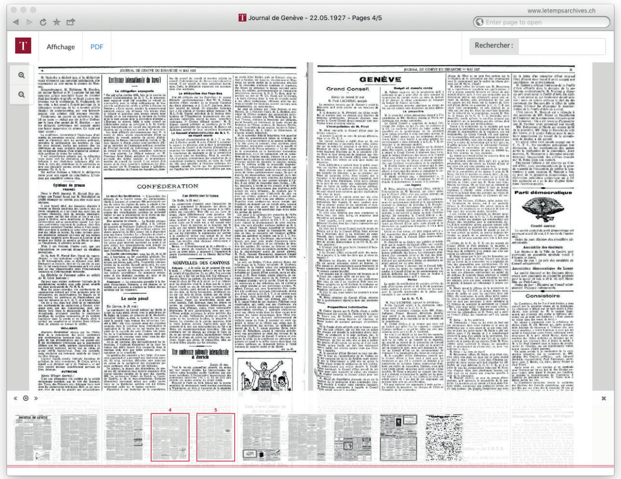
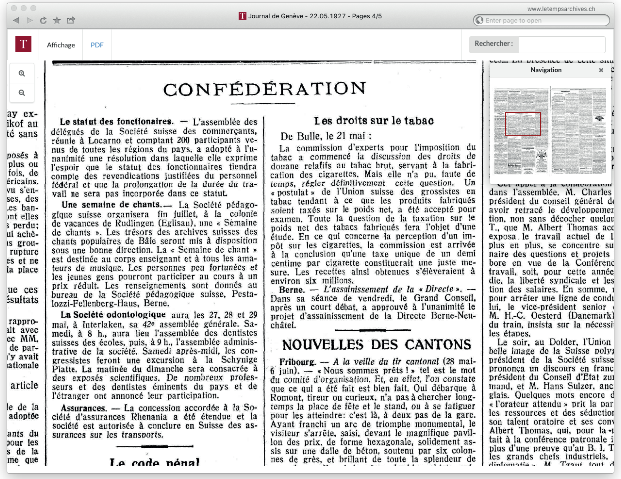


Figure 8: Viewer interface (featuring one full double-page). At the bottom, previews of all pages from the same issue.



CONFÉDÉRATION

Le statut des fonctionaires. — L'assemblée des délégués de la Société suisse des commerçants, réunie à Locarno et comptant 200 participants venus de toutes les régions du pays, a adopté à l'unanimité une résolution dans laquelle elle exprime l'espoir que le statut des fonctionnaires tiendra compte des revendications justifiées du personnel fédéral et que la prolongation de la durée du travail ne sera pas incorporée dans ce statut.

Une semaine de chants.— La Société pédagogique suisse organisera fin juillet, à la colonie de vacances de Rudlingen (Eglisau), une « Semaine de chants ». Les trésors des archives suisses des chants populaires de Bâle seront mis à disposition sous une bonne direction. La « Semaine de chant » est destinée au corps enseignant et à tous les amateurs de musique. Les personnes peu fortunées et les jeunes gens pourront participer au cours à un prix réduit. Les renseignements sont donnés au bureau de la Société pédagogique suisse, Pestalozzi-Fellenberg-Haus, Berne.

La Société odontologique aura les 27, 28 et 29 mai, à Interlaken, sa 42e assemblée générale. Samedi, à 8 h., aura lieu l'assemblée des dentistes suisses des écoles, puis, à 9 h., l'assemblée administrative de la société. Samedi après-midi, les congressistes feront une excursion à la Schynige Platte. La matinée du dimanche sera consacrée à des exposés scientifiques. De nombreux professeurs et des dentistes éminents du pays et de l'étranger ont annoncé leur participation.

Assurances. — La concession accordée à la Société d'assurances Rhenania a été étendue et la société est autorisée à conclure en Suisse des assurances sur les transports.

Le code pénal

Les droits sur le tabac

De Bulle, le 21 mai :

La commission d'experts pour l'imposition du tabac a commencé la discussion des droits de douane relatifs au tabac brut, servant à la fabrication des cigarettes. Mais elle n'a pu, faute de temps, régler définitivement cette question. Un « postulat » de l'Union suisse des grossistes en tabac tendant à ce que les produits fabriqués soient taxés sur le poids net, a été accepté pour examen. Toute la question de la taxation sur le poids net des tabacs fabriqués fera l'objet d'une étude. En ce qui concerne la perception d'un impôt sur les cigarettes, la commission est arrivée à la conclusion qu'une taxe unique de un demi centime par cigarette constituerait une juste mesure. Les recettes ainsi obtenues s'élèveraient à environ six millions.

Berne. — *L'assainissement de la « Directe ».* — Dans sa séance de vendredi, le Grand Conseil, après un court débat, a approuvé à l'unanimité le projet d'assainissement de la Directe Berne-Neuchâtel.

NOUVELLES DES CANTONS

Fribourg. — *A la veille du tir cantonal* (28 mai-6 juin). — « Nous sommes prêts ! » tel est le mot du comité d'organisation. Et, en effet, l'on constate que ce qui a été fait est bien fait. Qui débarque à Romont, tireur ou curieux, n'a pas à chercher longtemps la place de fête et le stand, ou à se fatiguer pour les atteindre: c'est là, à deux pas de la gare. Ayant franchi un arc de triomphe monumental, le visiteur s'arrête, saisi, devant le magnifique pavillon des prix, de forme hexagonale, solidement assis sur une dalle de béton, soutenu par six colonnes de grès, et brillant de toute la splendeur de

Figure 9: Viewer interface (zooming on one article). In the top right corner, the location of the article in the double page is highlighted.

The named entity query engine features a simple interface, in the form of a SPARQL endpoint, showed in figure 10. In order to make it more accessible to non-technical users, we included 5 sample queries that can be tried out with a simple click.

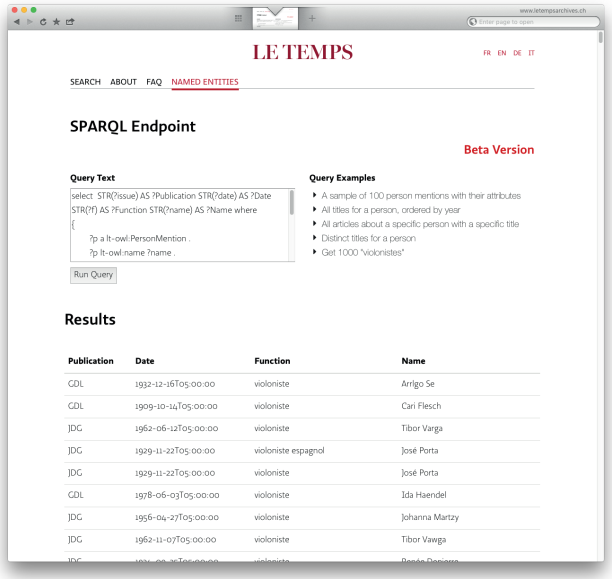
5.2 Application Stack Design

The software setup we decided on is as follows :

— Raw text indexation and search : Apache Solr.[24]

— Image Server : Loris IIIF.[25]

— Web development frameworks : Laravel.[26]



SPARQL Endpoint

Beta Version

Query Text

```
select  STR(?issue) AS ?Publication STR(?date) AS ?Date
STR(?f) AS ?Function STR(?name) AS ?Name where
{
    ?p a lt-owl:PersonMention .
    ?p lt-owl:name ?name .
```

Query Examples

- A sample of 100 person mentions with their attributes
- All titles for a person, ordered by year
- All articles about a specific person with a specific title
- Distinct titles for a person
- Get 1000 "violonistes"

Results

Publication	Date	Function	Name
GDL	1932-12-16T05:00:00	violoniste	Arrlgo Se
GDL	1909-10-14T05:00:00	violoniste	Cari Flesch
JDG	1962-06-12T05:00:00	violoniste	Tibor Varga
JDG	1929-11-22T05:00:00	violoniste espagnol	José Porta
JDG	1929-11-22T05:00:00	violoniste	José Porta
GDL	1978-06-03T05:00:00	violoniste	Ida Haendel
JDG	1956-04-27T05:00:00	violoniste	Johanna Martzy
JDG	1962-11-07T05:00:00	violoniste	Tibor Vawga

Figure 10: SPARQL endpoint presenting the results of a sample query.

— Internal database engine : PostgreSQL.[27]

— Triplestore : Virtuoso Open Source.[28]

Figure 11 shows the organisation of the different components. The typical web client issues a search request (*A*) that the web application forwards to the search engine (*B*) to find out the relevant pages, and to the internal database to load the necessary metadatas (*C*). Alternatively (in response to a SPARQL query), it will load data from the triplestore database (D). It then returns an HTML page (*E*) including URLs to the images that will be provided by the image server (*F*). Finally, new journal issues may be added to the archive using a publication workflow (*G*) that extracts image and textual representations from the scans.

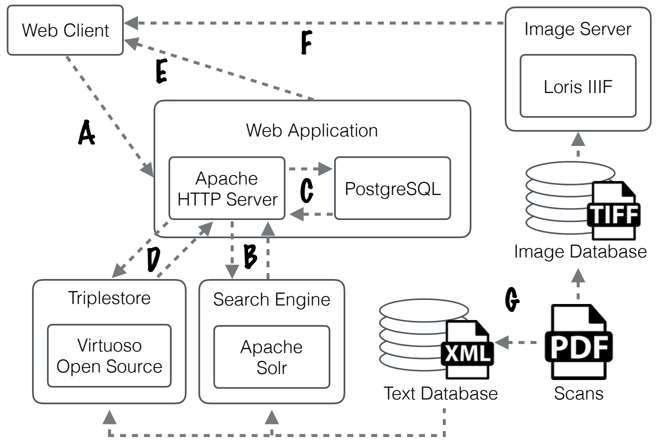


Figure 11: Information workflow and technical components.

5.3 Public Release

The final application has been released to the public on a dedicated web server running the full software stack described earlier. For improved performance, the website is cached using *Cloudflare* services.[29]

During the first month (March 18th to April 20th, 2016), about 35'000 search queries were made (hence more than 1'000 a day). Out of those, 2200 were direct accesses to specific dates, and the rest represented 21'350 unique words. According to Google Analytics[30], the new site was seen by 18'300 people, out of which more than 90% accessed it at least twice.

6. CONCLUSIONS

The new website and tools immediately received significant interest from researchers of several Swiss universities and state libraries. We received many constructive feedback, and answered questions from users having long-time use cases they needed to reproduce with the new web application.

Consultation statistics demonstrated great enthusiasm from the general public. On the day of the public launch, *Le Temps* published a dedicated article[31] and included a four-pages insert mainly composed of archival articles. Building on the launch, third parties also opened a *Facebook* page[32] to discuss noteworthy findings in the archives such as century old discussions relevant to current events or advertisements seen as comical from today's perspective.

Future works will focus on updating the contents and refining our tools to provide access to a wider range and even more relevant data depending on queries from users. Several improvement techniques have already been considered and are on their way :

— Improvement of raw data with a set of tools aiming to correct the OCR results, especially for the earlier years. Multiple approaches are possible including the use of language models [3], semi-automated statistical correction and crowdsourcing [14].

— Named entity disambiguation. In use, this allows the user to filter results that relate to different entities sharing the same names.

— Completion of the corpus with the missing journal issues, wherever possible.

— Find new partners and add new collections.

7. ACKNOWLEDGEMENTS

The authors would like to thank the Swiss National Library for their funding of this research, *Le Temps* for sharing their archives, as well as the Scientific Committee composed of Alain Clavien, Marie-Christine Doffey, Gaël Hurlimann, Joëlle Kuntz, Enrico Natale and François Vallotton, and the Geneva Library for their availability. All the actors cited provided precious and constructive feedbacks to our work.

[23] A double-page typically weighs about 10 MB.

[24] http://lucene.apache.org/solr

[25] http://github.com/loris-imageserver/loris

[26] http://laravel.com

[27] http://www.postgresql.org

[28] http://github.com/openlink/virtuoso-opensource

[29] http://www.cloudflare.com

[30] http://analytics.google.com

[31] http://www.letemps.ch/suisse/2016/03/18/epfl-temps-lancent-un-site-pointe-technologie-faciliter-acces-200-ans-archives Accessed on April 24th, 2016.

[32] http://www.facebook.com/groups/Etonnantdansletemps/ Accessed on April 24th, 2016.

8. REFERENCES

[1] S. Attfield and J. Dowell. Information seeking and use by newspaper journalists. *Journal of Documentation*, 59(2) :187–204, Apr. 2003.

[2] S. Auer, J. Lehmann, A.-C. Ngonga Ngomo, and A. Zaveri. Introduction to Linked Data and Its Lifecycle on the Web. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, S. Rudolph, G. Gottlob, I. Horrocks, and F. van Harmelen, editors, *Reasoning Web. Semantic Technologies for Intelligent Data Access*, pages 1–90. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[3] A. Bhardwaj, F. Farooq, H. Cao, and V. Govindaraju. Topic based language models for OCR correction. In *Proceedings of the SIGIR 2008 Workshop on Analytics for Noisy Unstructured Text Data*, pages 107–112. ACM Press, 2008.

[4] A. Bingham. 'The Digitization of Newspaper Archives : Opportunities and Challenges for Historians'. *Twentieth Century British History*, 21(2) :225–231, June 2010.

[5] V. Buntinx, C. Bornet, and F. Kaplan. Studying linguistic changes on 200 years of newspapers. In *DH2016 - Annual Conference of the Alliance of Digital Humanities Organizations*, 2016.

[6] V. Buntinx and F. Kaplan. Inversed N-gram viewer : Searching the space of word temporal profiles. In *DH2015 - Annual Conference of the Alliance of Digital Humanities Organizations*, 2015.

[7] K. Byrne. Nested named entity recognition in historical archive text. In *Semantic Computing, 2007. ICSC 2007. International Conference on*, pages 589–596. IEEE, 2007.

[8] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes. Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the* 9th *International Conference on Semantic Systems*, pages 121–124. ACM, 2013.

[9] M. Daniels and E. Yakel. Seek and You May Find : Successful Search in Online Finding Aid Systems. *The American Archivist*, 73(2) :535–568, Sept. 2010.

[10] M. Ehrmann, D. Nouvel, and S. Rosset. Named Entities Resources - Overview and Outlook. In *Proceedings of the* 10th *International Conference on Language Resources and Evaluation (LREC'10)*, Portorož, Slovenia, 2016 (to appear).

[11] F. Frontini, C. Brando, and J.-G. Ganascia. Semantic Web based Named Entity Linking for digital humanities and heritage texts. pages 1–12, 2015.

[12] R. Grishman and B. Sundheim. Design of the MUC-6 evaluation. In *Sixth Message Understanding Conference (MUC-6) : Proceedings of a Conference Held in Columbia, Maryland*, 1995.

[13] T. Heath and C. Bizer. Linked data : Evolving the web into a global data space. *Synthesis lectures on the semantic web : theory and technology*, 1(1) :1–136, 2011.

[14] R. Holley. Crowdsourcing : how and why should libraries do it ? *D-Lib Magazine*, 16(3) :4, 2010.

[15] L. James-Gilboe. The challenge of digitization : Libraries are finding that newspaper projects are not for the faint of heart. *The Serials Librarian*, 49(1-2) :155–163, 2005.

[16] E. Klijn. The current state-of-art in newspaper digitization : A market perspective. *D-Lib Magazine*, 14(1) :5, 2008.

[17] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight : shedding light on the web of documents. In *Proceedings of the* 7th *International Conference on Semantic Systems*, pages 1–8. ACM, 2011.

[18] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, , J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 2010.

[19] L. Monnet. La gazette de lausanne à l'origine. *Le Conteur vaudois*, 39(32) :1, 1901.

[20] A. Moro, A. Raganato, and R. Navigli. Entity Linking meets Word Sense Disambiguation : a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2 :231–244, 2014.

[21] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1) :3–26, 2007.

[22] C. Neudecker and A. Antonacopoulos. Making Europe's historical newspapers searchable. In *Proceedings of the 12th IAPR International Workshop on Document Analysis Systems (DAS2016)*, 2016.

[23] M. Perc. Evolution of the most common English words and phrases over the centuries. *Journal of The Royal Society Interface*, 9(77) :3323–3328, Dec. 2012.

[24] J. Piskorski. ExPRESS – Extraction Pattern Recognition Engine and Specification Suite. In *Proceedings of the International Workshop Finite-State Methods and Natural Language Processing 2007 (FSMNLP 2007)*, Potsdam, Germany, September 2007.

[25] K. Rodriquez, M. Bryant, T. Blanke, and M. Luszczynska. Comparison of named entity recognition tools for raw OCR text. In *KONVENS*, pages 410–414, 2012.

[26] S. Rosset, C. Grouin, K. Fort, O. Galibert, J. Kahn, and P. Zweigenbaum. Structured named entities in two distinct press corpora : Contemporary broadcast news and old newspapers. In *Proceedings. of the* 6th *Linguistic Annotation Workshop*, pages 40–48. Association for Computational Linguistics, 2012.

[27] R. S. Snydman, Stuart and T. Cramer. The international image interoperability framework (iiif) : A community & technology approach for web-based images. *Stanford University Libraries staff publications and research*, 2015.

[28] R. Steinberger, B. Pouliquen, M. Kabadjov, and E. van der Goot. JRC-Names : A Freely Available, Highly Multilingual Named Entity Resource. In *Proc. of the* 8th *International Conference Recent Advances in Natural Language Processing (RANLP'2011)*, Hissar, Bulgaria, September 2011.

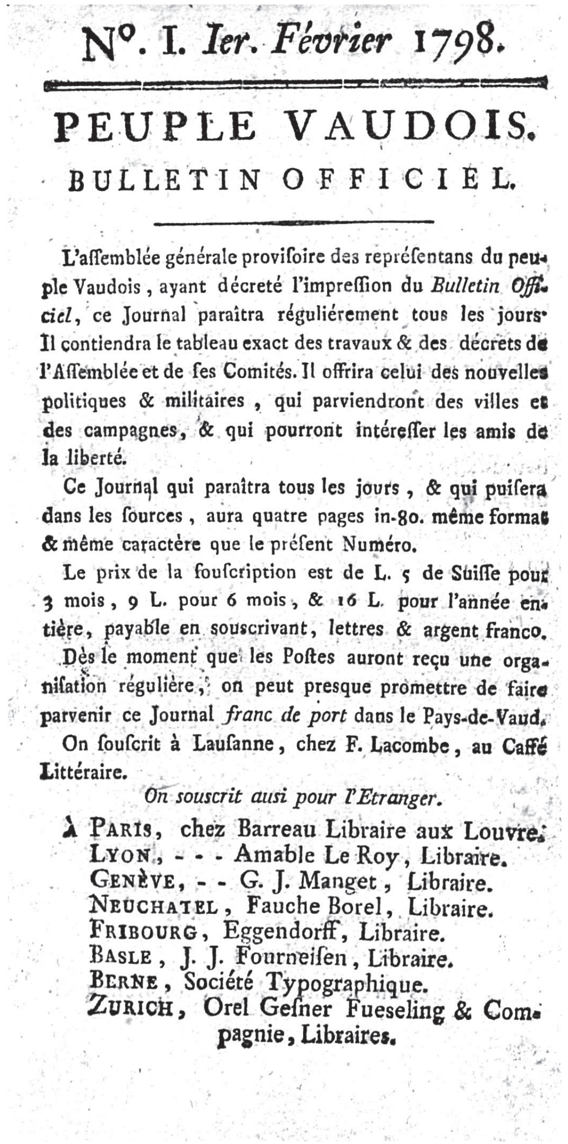
[29] R. Steinberger, B. Pouliquen, and E. van der Goot. An introduction to the europe media monitor family of applications. In . J. K. F. Gey, N. Kando, editor, *Information access in a multilingual world — Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR)*, Boston, USA, July 2009.

[30] S. van Hooland, M. De Wilde, R. Verborgh, T. Steiner, and R. Van de Walle. Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities*, 30(2) :262–279, 2015.

[31] E. Yakel. Encoded archival description : Are finding aids boundary spanners or barriers for users ? *Journal of Archival Organization*, 2(1-2) :63–77, June 2004.

[32] M. Yosef, J. Hoffart, I. Bordino, M. Spaniol, and G. Weikum. Aida : An online tool for accurate disambiguation of named entities in text and tables. In *Proceedings of the* 37th *International Conference on Very Large Databases* 9th, page 1450–1453, Seattle, USA, 2011.

APPENDIX



Nº. I. *Ier. Février* 1798.

PEUPLE VAUDOIS.

BULLETIN OFFICIEL.

L'aſſemblée générale proviſoire des repréſentans du peuple Vaudois, ayant décreté l'impreſſion du *Bulletin Officiel*, ce Journal paraîtra réguliérement tous les jours. Il contiendra le tableau exact des travaux & des décrets de l'Aſſemblée et de ſes Comités. Il offrira celui des nouvelles politiques & militaires, qui parviendront des villes et des campagnes, & qui pourront intéreſſer les amis de la liberté.

Ce Journal qui paraîtra tous les jours, & qui puiſera dans les ſources, aura quatre pages in-8o. même format & même caractère que le préſent Numéro.

Le prix de la ſouſcription est de L. 5 de Suiſſe pour 3 mois, 9 L. pour 6 mois, & 16 L. pour l'année entière, payable en souscrivant, lettres & argent franco.

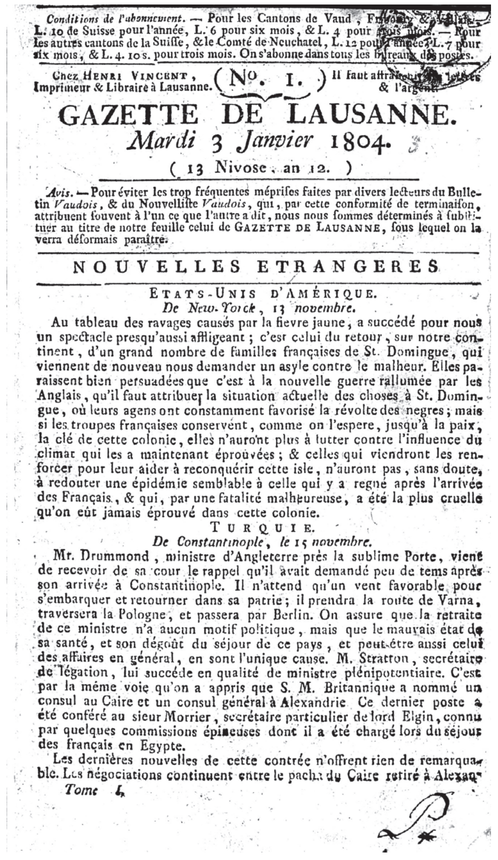
Dès le moment que les Poſtes auront reçu une organiſation régulière, on peut presque promettre de faire parvenir ce Journal *franc de port* dans le Pays-de-Vaud.

On ſouſcrit à Lauſanne, chez F. Lacombe, au Caffé Littéraire.

On souscrit ausi pour l'Etranger.

À PARIS, chez Barreau Libraire aux Louvre.
LYON, - - - Amable Le Roy, Libraire.
GENÈVE, - - G. J. Manget, Libraire.
NEUCHATEL, Fauche Borel, Libraire.
FRIBOURG, Eggendorff, Libraire.
BASLE, J. J. Fourneiſen, Libraire.
BERNE, Société Typographique.
ZURICH, Orel Geſner Fueseling & Compagnie, Libraires.

Figure 12: On February 1st, 1798, the front page of the first issue of what would later become *Gazette de Lausanne* after bearing nine other names [19]. "Bulletin officiel" approximately means "Official news report". This page shows letters s printed as f's.



Conditions de l'abonnement. — Pour les Cantons de Vaud, Fribourg & Valais, L. 10 de Suiſſe pour l'année, L. 6 pour six mois, & L. 4 pour trois mois. — Pour les autres cantons de la Suiſſe, & le Comté de Neuchatel, L. 12 pour l'année, L. 7 pour six mois, & L. 4. 10 s. pour trois mois. On s'abonne dans tous les bureaux des postes.

Chez Henri Vincent, Imprimeur & Libraire à Lausanne. (No. 1.) Il faut affranchir les lettres & l'argent.

GAZETTE DE LAUSANNE.

Mardi 3 *Janvier* 1804.

(13 Nivose, an 12.)

Avis. — Pour éviter les trop fréquentes mépriſes faites par divers lecteurs du Bulletin *Vaudois*, & du Nouvelliſte *Vaudois*, qui, par cette conformité de terminaiſon, attribuent ſouvent à l'un ce que l'autre a dit, nous nous ſommes déterminés à ſubſtituer au titre de notre feuille celui de Gazette de Lausanne, ſous lequel on la verra déſormais paraître.

NOUVELLES ETRANGERES

ETATS-UNIS D'AMÉRIQUE.

De New-Yorck, 13 *novembre*.

Au tableau des ravages causés par la fievre jaune, a succédé pour nous un spectacle presqu'aussi affligeant; c'est celui du retour, sur notre continent, d'un grand nombre de familles françaises de St. Domingue, qui viennent de nouveau nous demander un asyle contre le malheur. Elles paraissent bien persuadées que c'est à la nouvelle guerre rallumée par les Anglais, qu'il faut attribuer la situation actuelle des choses à St. Domingue, où leurs agens ont constamment favorisé la révolte des negres; mais si les troupes françaises conservent, comme on l'espere, jusqu'à la paix, la clé de cette colonie, elles n'auront plus à lutter contre l'influence du climat qui les a maintenant éprouvées; & celles qui viendront les renforcer pour leur aider à reconquérir cette isle, n'auront pas, sans doute, à redouter une épidémie semblable à celle qui y a regné après l'arrivée des Français, & qui, par une fatalité malheureuse, a été la plus cruelle qu'on eût jamais éprouvé dans cette colonie.

TURQUIE.

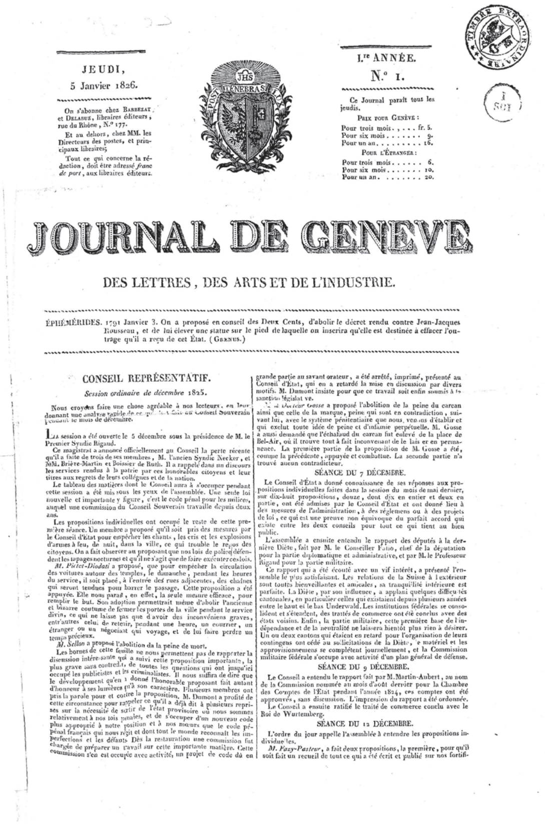
De Constantinople, le 15 *novembre*.

Mr. Drummond, ministre d'Angleterre près la sublime Porte, vient de recevoir de sa cour le rappel qu'il avait demandé peu de tems après son arrivée à Constantinople. Il n'attend qu'un vent favorable pour s'embarquer et retourner dans sa patrie; il prendra la route de Varna, traversera la Pologne, et passera par Berlin. On assure que la retraite de ce ministre n'a aucun motif politique, mais que le mauvais état de sa santé, et son dégoût du séjour de ce pays, et peut-être aussi celui des affaires en général, en sont l'unique cause. M. Stratton, secrétaire de légation, lui succéde en qualité de ministre plénipotentiaire. C'est par la même voie qu'on a appris que S. M. Britannique a nommé un consul au Caire et un consul général à Alexandrie Ce dernier poste a été conféré au sieur Morrier, secrétaire particulier de lord Elgin, connu par quelques commissions épineuses dont il a été chargé lors du séjour des français en Egypte.

Les dernières nouvelles de cette contrée n'offrent rien de remarquable. Les négociations continuent entre le pacha du Caire retiré à Alexan-

Tome I.

Figure 13: On January 3rd, 1804, few years after its creation, *Gazette de Lausanne* receives a name it kept for close to two centuries.



JEUDI,

5 Janvier 1826.

On s'abonne chez Barbezat et Delarue, libraires éditeurs, rue du Rhône, N.° 177.

Et au dehors, chez MM. les Directeurs des postes, et principaux libraires;

Tout ce qui concerne la rédaction, doit être adressé *franc de port*, aux libraires éditeurs.

JHS

POST TENEBRAS LUX

I.re ANNÉE.

N.° 1.

Ce Journal paraît tous les jeudis.

Prix pour Genève:

Pour trois mois . . . . fr. 5.
Pour six mois . . . . 9.
Pour un an . . . . 16.

Pour l'Étranger:

Pour trois mois . . . . . . 6.
Pour six mois . . . . 10.
Pour un an. . . . . 20.

TIMBRE EXTRAORDINAIRE

JOURNAL DE GENEVE

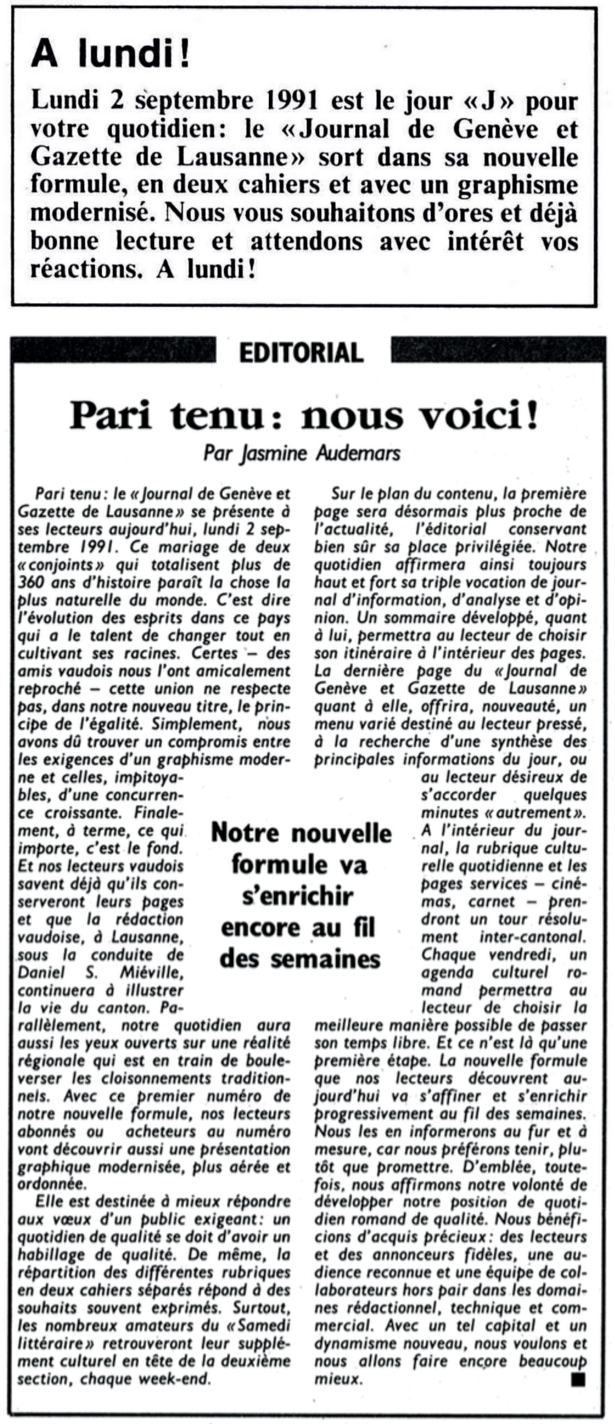
DES LETTRES, DES ARTS ET DE L'INDUSTRIE.

ÉPHÉMÉRIDES. 1791 Janvier 3. On a proposé en conseil des Deux Cents, d'abolir le décret rendu contre Jean-Jacques Rousseau, et de lui élever une statue sur le pied de laquelle on inscrira qu'elle est destinée à effacer l'outrage qu'il a reçu de cet État. (Grenus.)

CONSEIL REPRÉSENTATIF.

Session ordinaire de décembre 1825.

Figure 14: *Journal de Genève* was launched on January 5th, 1826.



A lundi!

Lundi 2 septembre 1991 est le jour «J» pour votre quotidien: le «Journal de Genève et Gazette de Lausanne» sort dans sa nouvelle formule, en deux cahiers et avec un graphisme modernisé. Nous vous souhaitons d'ores et déjà bonne lecture et attendons avec intérêt vos réactions. A lundi!

EDITORIAL

Pari tenu: nous voici!

Par Jasmine Audemars

Pari tenu: le «Journal de Genève et Gazette de Lausanne» se présente à ses lecteurs aujourd'hui, lundi 2 septembre 1991. Ce mariage de deux «conjoints» qui totalisent plus de 360 ans d'histoire paraît la chose la plus naturelle du monde. C'est dire l'évolution des esprits dans ce pays qui a le talent de changer tout en cultivant ses racines. Certes – des amis vaudois nous l'ont amicalement reproché – cette union ne respecte pas, dans notre nouveau titre, le principe de l'égalité. Simplement, nous avons dû trouver un compromis entre les exigences d'un graphisme moderne et celles, impitoyables, d'une concurrence croissante. Finalement, à terme, ce qui importe, c'est le fond. Et nos lecteurs vaudois savent déjà qu'ils conserveront leurs pages et que la rédaction vaudoise, à Lausanne, sous la conduite de Daniel S. Miéville, continuera à illustrer la vie du canton. Parallèlement, notre quotidien aura aussi les yeux ouverts sur une réalité régionale qui est en train de bouleverser les cloisonnements traditionnels. Avec ce premier numéro de notre nouvelle formule, nos lecteurs abonnés ou acheteurs au numéro vont découvrir aussi une présentation graphique modernisée, plus aérée et ordonnée.

Elle est destinée à mieux répondre aux vœux d'un public exigeant: un quotidien de qualité se doit d'avoir un habillage de qualité. De même, la répartition des différentes rubriques en deux cahiers séparés répond à des souhaits souvent exprimés. Surtout, les nombreux amateurs du «Samedi littéraire» retrouveront leur supplément culturel en tête de la deuxième section, chaque week-end.

Sur le plan du contenu, la première page sera désormais plus proche de l'actualité, l'éditorial conservant bien sûr sa place privilégiée. Notre quotidien affirmera ainsi toujours haut et fort sa triple vocation de journal d'information, d'analyse et d'opinion. Un sommaire développé, quant à lui, permettra au lecteur de choisir son itinéraire à l'intérieur des pages. La dernière page du «Journal de Genève et Gazette de Lausanne» quant à elle, offrira, nouveauté, un menu varié destiné au lecteur pressé, à la recherche d'une synthèse des principales informations du jour, ou au lecteur désireux de s'accorder quelques minutes «autrement». A l'intérieur du journal, la rubrique culturelle quotidienne et les pages services – cinémas, carnet – prendront un tour résolument inter-cantonal. Chaque vendredi, un agenda culturel romand permettra au lecteur de choisir la meilleure manière possible de passer son temps libre. Et ce n'est là qu'une première étape. La nouvelle formule que nos lecteurs découvrent aujourd'hui va s'affiner et s'enrichir progressivement au fil des semaines. Nous les en informerons au fur et à mesure, car nous préférons tenir, plutôt que promettre. D'emblée, toutefois, nous affirmons notre volonté de développer notre position de quotidien romand de qualité. Nous bénéficions d'acquis précieux: des lecteurs et des annonceurs fidèles, une audience reconnue et une équipe de collaborateurs hors pair dans les domaines rédactionnel, technique et commercial. Avec un tel capital et un dynamisme nouveau, nous voulons et nous allons faire encore beaucoup mieux. ■

Notre nouvelle formule va s'enrichir encore au fil des semaines

Figure 15: (Top.) On August 31st, 1991, a discreet insert at the bottom right corner of the front page announces that the two newspapers are merged into a single one. (Bottom.) On September 2nd, 1991, the result of the merged newspapers is published under the name *Journal de Genève et Gazette de Lausanne*.

JA 1211 GENÈVE 11

Samedi 28 février, dimanche 1er mars 1998
No 48 - 2 FS (tva 2% incluse) 10 FF

JOURNAL DE GENÈVE

et Gazette de Lausanne

Fondé en 1826 LE QUOTIDIEN SUISSE D'AUDIENCE INTERNATIONALE

LE DERNIER NUMÉRO

ÉDITORIAL

Au revoir

Par Ignace Jeannerat

Voici venu le jour où, pour la dernière fois, abonnés, lecteurs réguliers ou occasionnels, vos mains d'abord, puis votre regard se posent sur votre quotidien. Le titre de Une, l'Editorial, les fameuses «24 heures en deux minutes», etc.: instants magiques d'un échange répété à des milliers de reprises.

Ne pas avouer qu'à l'instant notre gorge est nouée serait mentir. On ne referme pas sans émotion l'histoire, la longue histoire de ce quotidien. Au point de citer cette réplique de Racine dans *Bérénice*: «Et pour jamais, adieu. Pour jamais! Ah, Seigneur, songez en vous-même combien ce mot cruel est affreux quand on aime!»

Une journée pour en finir

Vendredi 27 février 1998: le jour où le *Journal de Genève et Gazette de Lausanne* a disparu comme il a vécu.

Six heures trente. Quatre ombres se dirigent vers le siège de la Société suisse de radiodiffusion. Pour le rédacteur en chef Ignace Jeannerat, le photographe Daniel Winteregg et les journalistes Luis Lema et Sylvain Besson, le dernier jour du *Journal de Genève et Gazette de Lausanne* commence par un horaire de condamnés à mort. Avant l'aube, dans la nuit qui précède une journée qui s'annonce belle quoique fraîchissante vers le soir.

Vendredi 27 février, 7 heures du matin, dans les studios de la Radio romande: le chef de la rubrique internationale, Luis Lema (à g.), et Sylvain Besson, rédacteur à la rubrique suisse, entourent Marlène Belilos, qui anime une émission de France Culture consacrée à la fin du «Journal».

Figure 16: February 28th, 1998. The final issue of *Journal de Genève et Gazette de Lausanne*. It would then be merged with *Le Nouveau Quotidien* in order to form *Le Temps*, which was first issued on March 18th, 1998.

Towards a Systematic Information Governance Maturity Assessment

Diogo Proença
IST / INESC-ID
Rua Alves Redol, 9
1000-029 Lisboa
diogo.proenca@tecnico.ulisboa.pt

Ricardo Vieira
IST / INESC-ID
Rua Alves Redol, 9
1000-029 Lisboa
rjcv@tecnico.ulisboa.pt

José Borbinha
IST / INESC-ID
Rua Alves Redol, 9
1000-029 Lisboa
jlb@tecnico.ulisboa.pt

ABSTRACT

Information Governance as defined by Gartner is the “specification of decision rights and an accountability framework to encourage desirable behavior in the valuation, creation, storage, use, archival and deletion of information. Includes the processes, roles, standards and metrics that ensure the effective and efficient use of information in enabling an organization to achieve its goals”. In this paper, we present assess the maturity of seven project pilots using the Information Governance maturity model based on existing reference documents. The process is based on existing maturity model development methods. These methods allow for a systematic approach to maturity model development backed up by a well-known and proved scientific research method called Design Science Research. An assessment was conducted and the results are presented in this paper, this assessment was conducted as a self-assessment in the context of the EC-funded E-ARK project for the seven pilots of the project. The main conclusion from this initial assessment is that there is much room for improvement with most pilots achieving results between maturity level two and three. As future work, the goal is to analyze other references from different domains, such as, records management. These references will enhance, detail and help develop the maturity model making it even more valuable for all types of organization that deal with information governance.

KEYWORDS

Information Governance, Maturity Assessment, Maturity Model.

1. INTRODUCTION

A Maturity Model consists of a number of entities, including “maturity levels” (often six) which are, from the lowest to the highest, (0) Non Existent, (1) Initial, (2) Basic, (3) Intermediate, (4) Advanced and (5) Optimizing. Each aspect can have its own Maturity Model, which expresses quantitatively the maturity level of an organization regarding a certain aspect. A Maturity Model provides also a way for organizations to see clearly what they must accomplish in order to pass to the next maturity level.

The use of maturity models is widespread and accepted, both in industry and academia. There are numerous maturity models, with at least one for each of the most trending topics in such areas as Information Technology or Information Systems. Maturity models are widely used and accepted because of their simplicity and effectiveness. They can help an organization to understand the current level of maturity of a certain aspect in a meaningful way, so that stakeholders can clearly identify strengths to be built upon and weaknesses requiring improvement, and thus prioritize what must be done in order to reach a higher level. This can be used to show the outcomes that will result from that effort, enabling stakeholders to decide if the outcomes justify the effort.

There are several examples of maturity models currently in use. For example, in software engineering there is the classic Software Engineering Institute Capability Maturity Model Integration also known as the CMMI that has been growing in the last twenty years, already covering a set of aspects regarding products and services lifecycles. In the Information Management domain there also several examples of maturity models such as the Gartner Enterprise Information Management Maturity Model. Other domains where maturity models can be found include management, business process management, energy management, governance and risk management, etc. The previous maturity models are already described and analyzed in [35], where a state of the art on maturity models was performed. We have also noted existing work in the area of a Digital Preservation Maturity Models undertaken by Adrian Brown where the author examines the notion of “trusted” digital repositories and proposes a maturity model for digital preservation, which goal is to enable organizations to assess their capabilities and create a roadmap for developing them to the required maturity level [8], and of Charles Dollar that proposes a Capability Maturity Model to assess digital preservations requirements [9] according to the Open Archival Information System (OAIS) Reference Model (ISO14721 [2]) and Trustworthy Repository Assessment Criteria (TRAC) Standard (ISO16363 [1]). Those maturity models will be analyzed in detail in E-ARK deliverable D7.5.

This paper builds on the knowledge from the maturity models that have been documented in detail in [35], process assessment and assessment in general and focus on assessing the maturity levels of the seven pilots of the E-ARK project:

- Pilot 1: SIP creation of relational databases (Danish National Archives);
- Pilot 2: SIP creation and ingest of records (National Archives of Norway);
- Pilot 3: Ingest from government agencies (National Archives of Estonia);
- Pilot 4: Business archives (National Archives of Estonia, Estonian Business Archives);
- Pilot 5: Preservation and access to records with geodata (National Archives of Slovenia);
- Pilot 6: Seamless integration between a live document management system and a long-term digital archiving and preservation service (KEEP SOLUTIONS);
- Pilot 7: Access to databases (National Archives of Hungary).

This paper is a continuation of the maturity development method presented in [35], and focuses on the three final steps of the development method which are detailed in Section 3. In Section 4 the self-assessment questionnaire used to perform the assessment is detailed. Then, in Section 5, the results of the assessment are detailed and analyzed. Section 6 details the post-assessment feedback questionnaire analysis and conclusions. Finally, Section 7 presents the conclusions of this paper.

2. RELATED WORK

This section details the related work relevant for this paper, namely the maturity model fundamentals and maturity assessment methods. These are essential to understand the remaining of this paper.

2.1 Maturity Model Fundamentals

To evaluate maturity, organizational assessment models are used, which are also known as stages-of-growth models, stage models, or stage theories [23].

The concept of maturity is a state in which, when optimized to a particular organizational context, is not advisable to proceed with any further action. It is not an end, because it is a mobile and dynamic goal [14]. It is rather a state in which, given certain conditions, it is agreed not to continue any further action. Several authors have defined maturity, however many of the current definitions fit into the context in which each a particular maturity model was developed.

In [15] maturity is defined as a specific process to explicitly define, manage, measure and control the evolutionary growth of an entity. In turn, in [16] maturity is defined as a state in which an organization is perfectly able to achieve the goals it sets itself. In [17] it is suggested that maturity is associated with an evaluation criterion or the state of being complete, perfect and ready and in [18] as being a concept which progresses from an initial state to a final state (which is more advanced), that is, higher levels of maturity. Similarly, in [19] maturity is related with the evolutionary progress in demonstrating a particular capacity or the pursuit of a certain goal, from an initial state to a final desirable state. Still, in [20] it is emphasized the fact that this state of perfection can be achieved in various ways. The distinction between organizations with more or less mature systems relates not only to the results of the indicators used, but also with the fact that mature organizations measure different indicators when comparing to organizations which are less mature [21]. While the concept of maturity relates to one or more items identified as relevant [22], the concept of capability is concerned only with each of these items. In [23] maturity models are defined as a series of sequential levels, which together form an anticipated or desired logical path from an initial state to a final state of maturity. These models have their origin in the area of quality [24][25]. The Organizational Project Management Maturity Model (OPM3) defines a maturity model as a structured set of elements that describe the characteristics of a process or product [26][27]. In [28] maturity models are defined as tools used to evaluate the maturity capabilities of certain elements and select the appropriate actions to bring the elements to a higher level of maturity. Conceptually, these represent stages of growth of a capability at qualitative or quantitative level of the element in growth, in order to evaluate their progress relative to the defined maturity levels.

Some definitions found involve organizational concepts commonly used, such as the definition of [29] in which the authors consider a maturity model as a "... a framework of evaluation that allows an organization to compare their projects and against the best practices or the practices of their competitors, while defining a structured path for improvement." This definition is deeply embedded in the concept of benchmarking. In other definitions, such as in the presented by [30] there appears the concern of associating a maturity model to the concept of continuous improvement.

In [31], the maturity models are particularly important for identifying strengths and weaknesses of the organizational context to which they are applied, and the collection of information through methodologies associated with benchmarking. In [32] it was concluded that the great advantage of maturity models is that they show that maturity must evolve through different dimensions and, once reached a maturity level, sometime is needed for it to be actually sustained. In [33] it was concluded that project performance in organizations with higher maturity levels was significantly increased. Currently, the lack of a generic and global standards for maturity models has been identified as the cause of poor dissemination of this concept.

2.2 Maturity Assessment

An assessment is a systematic method for obtaining feedback on the performance of an organization and identify issues that affect performance. Assessments are of extreme importance as organizations are constantly trying to adapt, survive, perform and influence despite not being always successful. To better understand what they can or should change to improve the way they conduct their business, organizations can perform organizational assessments. This technique can help organizations obtain data on their performance, identify important factors that help or inhibit the achievement of the desired outcomes of a process, and benchmark them in respect to other organizations. In the last decade, the demand for organizational assessment are gaining ground with the implementation of legislation that mandate good governance in organizations, such as, the Sarbanes-Oxley Act [7] and the BASEL accords in financial organizations [8]. Moreover, funding agencies are using the results of these assessments to understand the performance of organizations which they fund (e.g., Not for profit organizations, European Commission, Banks, Research institutes) as a means to determine how well organizations are developing the desired outcomes, and also to better understand the capabilities these organizations have in place to support the achievement of the desired outcome.

The result of an assessment effort will be a set of guidelines which will allow for process improvement. Process improvement is a way of improving the approach taken for organizing and managing business processes and can involve also executing improvements to existing systems. There are several examples of process improvement such as compliance with existing legislation. Process improvement often results in process redesign which involves understanding the requirements of a stakeholder and developing processes which meet the stakeholders’ expectations. This often means that the existing processes supporting a specific part of business need to be adapted, or even made from scratch to meet the stakeholders’ expectations. When the processes need to be made from scratch we are dealing with process reengineering which is a way to introduce radical changes in the business processes of an organization and changes the way a business operates. In this way, process reengineering starts from scratch by determining how the key business activities need to be reengineered to meet stakeholders’ expectations. One well known example, is the transition from traditional banking services to on-line banking services.

The ISO/IEC 15504, describes a method that can be used to guide the assessment of organizational processes, which is depicted in Figure 1. The ISO15504 assessment method is composed of seven main steps which are then further detailed in atomic tasks.

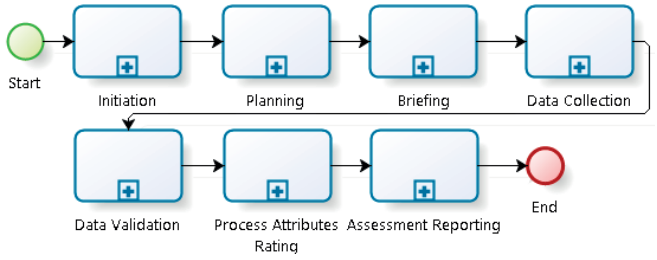


Figure 1. ISO15504 Assessment Process Overview.

3. ASSESSMENT PROCESS

One recurrent criticism of maturity models is that they lack empirical foundation and traceability [7]. The main reason for the criticism is that existing maturity models typically do not follow a theoretical framework or methodology for their development [7]. In fact, there is an absence on literature regarding methods and practices for the design and development of maturity models [7].

One of the most known development model for maturity models is the one from Becker in [4], a procedure based on a scientific research method called Design Science Research (DSR). The well-argued claim of the design procedure [4] is that these fundamental requirements should drive the development of every maturity model. Apart from evaluating well-known models according to these dimensions, the article also delineates a set of steps to correctly develop a maturity model. It depicts which documentation should result from each step, and includes an iterative maturity model development method that proposes that each iteration of the maturity model should be implemented and validated before going to a new iteration. The procedure delineates eight requirements [4], (1) Comparison with existing maturity models is presented and clearly argues for the need of a new model or the adaptation of an existing one; (2) Iterative Procedures are followed to ensure a feedback loop and refinement; (3) The principles, quality and effectiveness behind the design and development effort of a maturity model should pass through an iterative Evaluation step; (4) The design and development of maturity models should follow a Multi-methodological Procedure which use must be well founded; (5) During the development of a maturity model there should be a clear Identification of Problem Relevance so that the problem solution can be relevant to practitioners and researchers; (6) Problem Definition should include the application domain for the maturity model and also detail the intended benefits and constraints of application; (7) There should be a Targeted Presentation of Results regarding the users' needs and application constraints and, (8) The design of a maturity model must include Scientific Documentation, which details the whole process design for each step of the process, as well as, the methods applied, people involved and the obtained results.

One limitation of existing maturity models is that it is not typically not clear which requirements were used for the design and development of the model. In other words, there is a weak or inexistent traceability between the maturity model and the requirements that are used as reference. Consequently, stakeholders that wish to use the maturity model are unable to understand if the model is aligned with current best practices. To address the aforementioned traceability problem the maturity model described in this paper is based in well-known references of IG. Due to the fact that IG is a multi-disciplinary fields that covers several disciplines the range of standards and references documents is vast and include references, such as, the ISO 16363, ISO 20652, ISO 14721, MoREQ 2010, ISO 16175, ISO 23081, ISO 30301, ISO 27001, among others.

The maturity model for information governance, depicted further on in this section, consists of three dimensions:

- **Management:** "The term management refers to all the activities that are used to coordinate, direct, and control an organization." [12]
- **Processes:** "A process is a set of activities that are interrelated or that interact with one another. Processes use resources to transform inputs into outputs." [12]
- **Infrastructure:** "The term infrastructure refers to the entire system of facilities, equipment, and services that an organization needs in order to function."[12]

These dimensions provide different viewpoints of information governance which help to decompose the maturity model and enable easy understanding.

For each dimension we have a set of levels, from one to five, where one show the initial phase of maturity of a dimension and level five shows that the dimension is fully mature, self-aware and optimizing. These levels and their meaning were adapted from the levels defined for SEI CMMI. [13]

In order to assess the E-ARK pilots on their maturity regarding information governance, the project has adopted a self-assessment process. In this self-assessment process, a questionnaire is provided to the organization to be assessed which they complete to the best of their knowledge. Then the results are analysed by the assessment team and an assessment report is provided to the organization. This paper continues the application of the maturity model development method presented in [36] (and reproduced on Figure 2) and focuses on the application of the maturity model on the use cases before the project pilot, i.e. the three last stages of the method. E-ARK Deliverable 7.5 will use the results presented here to further develop and extend the maturity model. Finally, in E-ARK deliverable 7.6 will use the final maturity model to perform a final assessment of the project pilots.

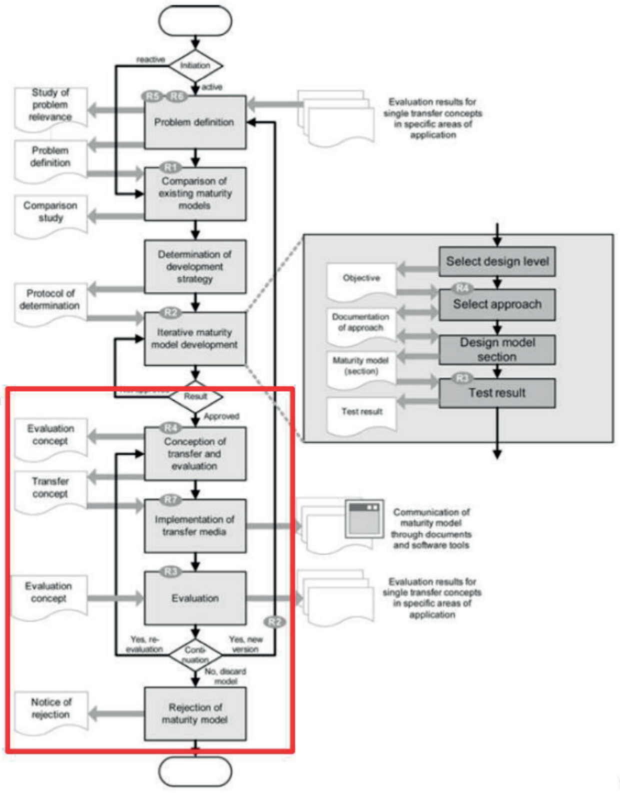

reactive
Initiation
active
R5
R6
Study of problem relevance
Problem definition
Problem definition
Evaluation results for single transfer concepts in specific areas of application
R1
Comparison of existing maturity models
Comparison study
Determination of development strategy
Protocol of determination
R2
Iterative maturity model development
Select design level
Objective
R4
Select approach
Documentation of approach
Design model section
Maturity model (section)
R3
Test result
Test result
Result
Approved
Evaluation concept
R4
Conception of transfer and evaluation
Transfer concept
R7
Implementation of transfer media
Communication of maturity model through documents and software tools
Evaluation concept
R3
Evaluation
Evaluation results for single transfer concepts in specific areas of application
R2
Yes, re-evaluation
Continuation
Yes, new version
No, discard model
Notice of rejection
Rejection of maturity model

Figure 2. Maturity Model Design Procedure [4]

The concept of transfer and evaluation of the maturity model was defined through the identification of the pilots' capabilities. A capability can be defined as "an ability that an organization, person, or system possesses" that typically requires a combination of "organization, people, processes, and technology" for its realization [3]. The definition of a capability must be implementation-independent, as it might be realized in different ways and measured in different levels of maturity.

Pilot's capabilities were identified through the analysis of [34] which details the E-ARK general pilot model and defines the purpose and processes of each pilot. Five top-level capabilities were defined: Pre-Ingest, Ingest, Archival Storage Preservation, Data Management, and Access. Table 1 depicts the defined capabilities and its corresponding abilities. As presented in the table, the pilots will have different focus and consequently will aim for different capabilities. For example, pilot 1 and 2 will focus merely on the capabilities of pre-ingest and ingest while other pilots contain the full lifecycle of pre-ingest, ingest, archival storage, data management and access.

The Pre-Ingest capability depicts the abilities to create submission information packages, encompassing the validation and enhancement of a SIP received from producers to create an E-ARK compliant SIP. The assessment of the maturity level must measure these abilities.

The Ingest capability reflects the abilities to create AIPs from the ingested SIPs. As most of the archival solutions available in the market make use of specific archival information packages, a high maturity level will include the creation of the E-ARK AIP from the E-ARK SIP. The Ingest capability also involves the ability to validate the E-ARK SIP received from pre-ingest.

The Archival Storage Preservation capability reflects the abilities to store and preserve the E-ARK AIP on the long term. As the focus of the project is particularly directed towards the processing phases surrounding the archival and preservation of data, the assessment will target the symbolic process of storing the E-ARK AIP.

The Data Management capability represents the ability to manipulate descriptive metadata, allowing the enhancement of existing E-ARK AIP, which will result in new E-ARK AIP.

Finally, the Access capability comprises the abilities to create the DIP, either on a local format or as E-ARK DIP, either on a pre-defined manner (defined as "standard" in the [34]), where the consumer accesses the requested data, or by special request producing a DIP in a local format or as E-ARK DIP, both produced using sophisticated analysis and presentation tools. An aspect to take into consideration, is that even though the pilots focus on a certain capability there might be abilities - a) to r) – that are not relevant in the context of a certain pilot and as result are no piloted.

Based on the capabilities definition the questionnaire was divided into five sections, which identify each capability:

(1) Pre-Ingest,
(2) Ingest,
(3) Archival Storage and Preservation,
(4) Data Management, and
(5) Access.

Using the defined capability model the assessment questionnaire was built by, for each ability, define one or more questions to assess the selected ability then, using the maturity model defined in [35], define the possible answers of the question(s).

The assessment of a particular capability will then evaluate the degree of realization and performance of the people, processes, and technology that comprise that capability.

One aspect to consider is that each question is created independent from all the others and all the questions have the same weight to the maturity level calculation. These questions are detailed in section 4.

4. SELF-ASSESSMENT QUESTIONNAIRE

This section details the self-assessment questionnaire used to assess the E-ARK pilots. The questionnaire is comprised of five capabilities which are detailed in the previous section, then each capability contains a set of questions. Each question is detailed in a table with the following fields:

1. **ID:** Which identifies the number of the question in the overall questionnaire;
2. **Title:** Which depicts the main topic the question refers to;
3. **Question:** Which details the question itself;
4. **Objective:** Which details the objective of that question, what knowledge the question intends to capture;
5. **Notes:** Which either clarifies some aspects and/or terms of the question or details examples of evidence to substantiate the answer for the question;

Table 1. Capability Model and the Pilots

Capability	Ability	Pilots						
		1	2	3	4	5	6	7
Pre-Ingest	a) SIP Content Definition b) Transformation of the Producer SIP to E-ARK SIP c) Local SIP Validation d) Enhancement of the local SIP e) Creation of the E-ARK SIP	F	F	F	F	F	F	F
Ingest	f) Creation of fonds g) Creation of the E-ARK AIP h) Validation of the E-ARK SIP i) Validation of the E-ARK AIP	F	F	F	F	T	F	F
Archival Storage and Preservation	j) Store E-ARK AIP		T	T	T	T	F	T
Data Management	k) Export E-ARK AIP and Descriptive metadata l) Enhance E-ARK AIP and Descriptive metadata			T	F	T	T	
Access	m)Search Data n) Provide Access to Ad-Hoc DIP o) Creation of a Local DIP p) Creation of a E-ARK DIP q) Creation of a Requested Local DIP r) Creation of a Requested E-ARK DIP	T		F	F	F	F	F

F	Focus of the pilot
T	Elements also used/tried within the pilot

6. **Terms:** Which identifies the terms that are detailed in EVOC. EVOC is the vocabulary manager which makes part of the knowledge centre being developed in E-ARK;
7. **Answers:** Which depicts the five possible answers to the question;
8. **Source:** Which details the source from which that specific question originates.

The questionnaire starts by providing an introduction. This introduction provides details on the purpose of the questionnaire, how it will be analysed, and clarifies concepts being constantly used throughout the questionnaire. [36] details the questionnaire that was presented to the respondents.

This questionnaire consists of a set of questions that will be used to determine the maturity level of the E-ARK pilots for each of the five capabilities of the E-ARK General Model. All questions are mandatory.

The answers provided will then be analysed by the Information Governance Maturity Model development team and a report will be issued detailing all the findings of the assessment. The set of assessment reports is available at [36].

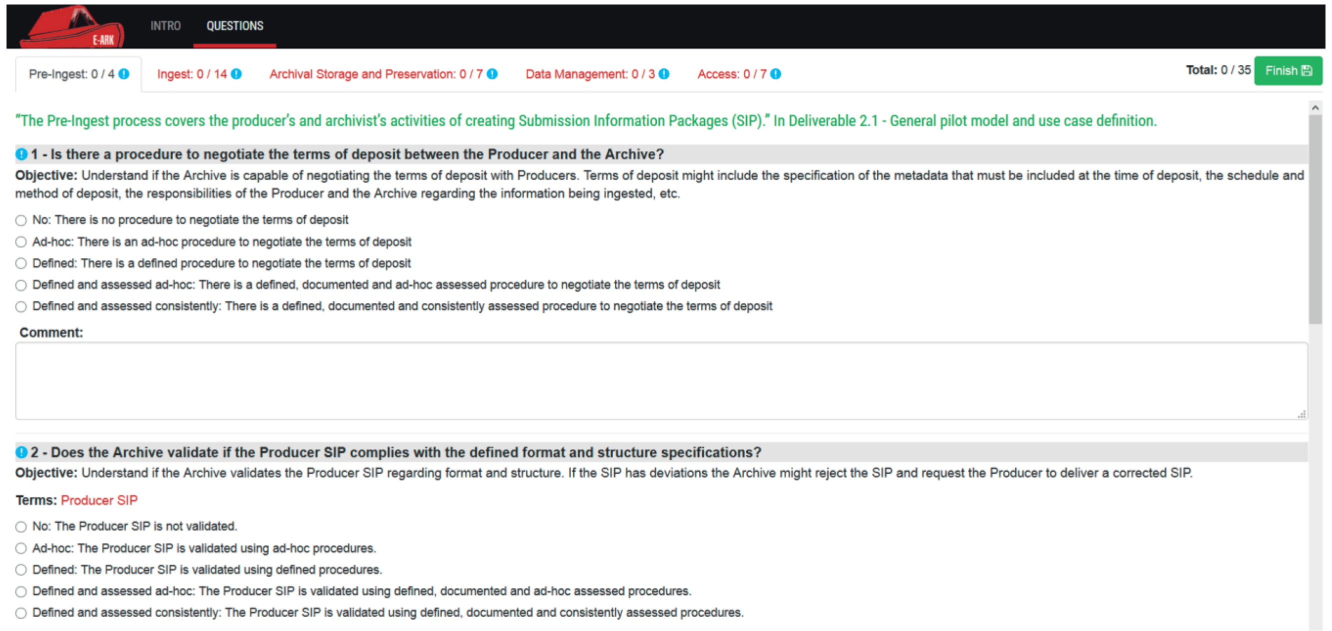
The questionnaire uses the following definitions of measurement:

- **No** indicates that there is no procedure or mechanism in place;
- **Ad-hoc** refers to actions performed but not being repeatable in the future, which can be due to the lack, outdate or no use of proper documentation, procedures or mechanisms, and thus leading to different people performing different tasks to achieve the same outcome;
- **Defined** refers the ways to achieve an outcome are supported by defined procedures or mechanisms, and thus leading to the actions performed being capable of being repeated in the future. This level does not give an assurance that the defined procedures or mechanisms are being consistently complied with or assessed;
- **Ad-hoc assessed** means that there is a concern with the assessment of some specific aspects, but that is not performed under a defined process but ad-hoc and when the need arises**;**
- **Consistently assessed** means that there is a concern with the assessment of some specific aspects, and that such is performed continuously, under a defined process, with alerts triggered by a defined set of indicators considering these dimensions, for example:
  - **Completeness**, which focuses on assessing if a procedure performs all relevant steps, aligned with the most recent documented requirements for that;
  - **Effectiveness**, which focus on assessing if the results of a procedure are free of errors and do not require further handling;
  - **Efficiency**, which focus on assessing if a procedure executes with the optimal efforts (for example, if automation is used instead of human effort), in an agreed time period as to avoid bottlenecks on the infrastructure and to minimize the time spent on executing it;
  - **Relevance**, which focus on assessing if the implemented requirements are still relevant for the intended purpose (as legislation change, for example, there is the need to assess if implemented requirements are still relevant).

These are just examples of aspects that need to be measured at higher levels of maturity, there might be further aspects to measure depending on the specific requirements of the pilot.

For each question there is a field respondents can use to provide additional comments, clarifications or a justification to the answer. These comments will be considered by the assessment team when evaluating the answers.

The questionnaire was sent to the pilot owners and was available on-line at http://earksurvey.sysresearch.org. The questionnaire was presented in a set of five tabs, one for each of the capabilities identified. Then in each tab a short description of the capability is presented followed by the questions, objective, notes, terms, answers and a field for comments (shown in Figure 3).



Pre-Ingest: 0 / 4 Ingest: 0 / 14 Archival Storage and Preservation: 0 / 7 Data Management: 0 / 3 Access: 0 / 7 **Total:** 0 / 35

"The Pre-Ingest process covers the producer's and archivist's activities of creating Submission Information Packages (SIP)." In Deliverable 2.1 - General pilot model and use case definition.

1 - Is there a procedure to negotiate the terms of deposit between the Producer and the Archive?

Objective: Understand if the Archive is capable of negotiating the terms of deposit with Producers. Terms of deposit might include the specification of the metadata that must be included at the time of deposit, the schedule and method of deposit, the responsibilities of the Producer and the Archive regarding the information being ingested, etc.

○ No: There is no procedure to negotiate the terms of deposit
○ Ad-hoc: There is an ad-hoc procedure to negotiate the terms of deposit
○ Defined: There is a defined procedure to negotiate the terms of deposit
○ Defined and assessed ad-hoc: There is a defined, documented and ad-hoc assessed procedure to negotiate the terms of deposit
○ Defined and assessed consistently: There is a defined, documented and consistently assessed procedure to negotiate the terms of deposit

Comment:

2 - Does the Archive validate if the Producer SIP complies with the defined format and structure specifications?

Objective: Understand if the Archive validates the Producer SIP regarding format and structure. If the SIP has deviations the Archive might reject the SIP and request the Producer to deliver a corrected SIP.

Terms: Producer SIP

○ No: The Producer SIP is not validated.
○ Ad-hoc: The Producer SIP is validated using ad-hoc procedures.
○ Defined: The Producer SIP is validated using defined procedures.
○ Defined and assessed ad-hoc: The Producer SIP is validated using defined, documented and ad-hoc assessed procedures.
○ Defined and assessed consistently: The Producer SIP is validated using defined, documented and consistently assessed procedures.

Figure 3. On-line Self-Assessment Questionaire

5. SELF-ASSESSMENT RESULT ANALYSIS

This section details the analysis of the results for each of the E-ARK pilots. For each pilot, in [36], the following is provided:

1. The answer provided for each question;
2. The comments provided in each question, in case there is a comment;
3. The weak points, aspects that should be considered for improvement;
4. The maturity level for each of the capabilities of the questionnaire.

It is important to note that for the purpose of this paper we are only assessing the "Processes" dimension of the Information Governance Maturity Model. This is due to the fact that the E-ARK pilots do not have an organizational background which would allow assessing the other two dimensions. The results are calculated as an average of the maturity levels of the questions for each capability, this average was then rounded down.

In the conclusion of this section there is a comparison and analysis between the pilots, regarding the findings of the self-assessment. Table 2 details the maturity levels of answers provided to each question by each pilot, as well as, the calculated maturity level for each of the capabilities of the questionnaire. For the result of each capability of each pilot there is an associated colour. This colour is linked to Table 1, where blue represents a focus capability and red a capability to be explored. The lack of these two colours means that that capability is not part of the pilot.

The answers provided will then be analysed by the Information Governance Maturity Model development team and a report will be issued detailing all the findings of the assessment. The set of assessment reports are available at http://www.eark-project.com/resources/project-deliverables/46-d72initassess/file.

Figure 4 depicts a comparison between the pilots. Pilot 1 is the one which achieved the best overall results, especially in pre-ingest and access it achieved the best results. Pilot 2 achieved the second best results. However there are still some enhancements to perform in the access capability where it achieved maturity level 2. Despite this fact, the access capability is not the focus in pilot 2. Pilot 7 also shows a high level maturity across the capabilities measured in the assessment. However, as in pilot 2, there are still some important enhancements to perform to the access capability. In pilot 7, the importance of the access capability is considerable due to it being one of the focuses of the pilot.

The other four pilots showed similar results among the capabilities. With some exceptions for pilot 3, where it shows higher maturity levels for pre-ingest and the access capabilities. Another exception is pilot 6 which shows higher maturity levels for ingest and data management capabilities. Pilot 5 did not answer to the questions for the archival storage and preservation and as the result no maturity level was calculated. As this is not the focus capability of the pilot there is no major issue with this fact.

There are still several capabilities at maturity level 1 or 2 for all pilots except pilot 1. These should be addressed as soon as possible to reach at least maturity level 3 for the focus capabilities. This is due to the fact that maturity level 3 is considered an intermediate level between lack of definition of consistency of mechanism and procedures typical of maturity level 1 and 2; and the documentation and assessment of mechanism and procedures typical of maturity level 4 and 5. Maturity level 3 depicts aspects that are consistent and defined throughout the organizational or pilot context and shows a state of change in this context from no definition to improvement. The outcomes of the E-ARK project will help the pilots to reach this maturity level and will also assist other organizations to reach higher levels of maturity and as result improve archival practice.

6. POST-ASSESSMENT FEEDBACK QUESTIONNAIRE

After analyzing and reporting the results of the initial assessment and evaluation, a post assessment questionnaire was developed. This questionnaire allowed pilots to provide feedback to the Information Governance Maturity Model Development Team to promote continuous improvement of the assessment process and the questionnaire used to assess the Information Governance Maturity Model.

For each question there was a three point answer scale, with possible answers of (1) Yes, (2) Partially and (3) No. For each question comments could be provided to detail the answers.

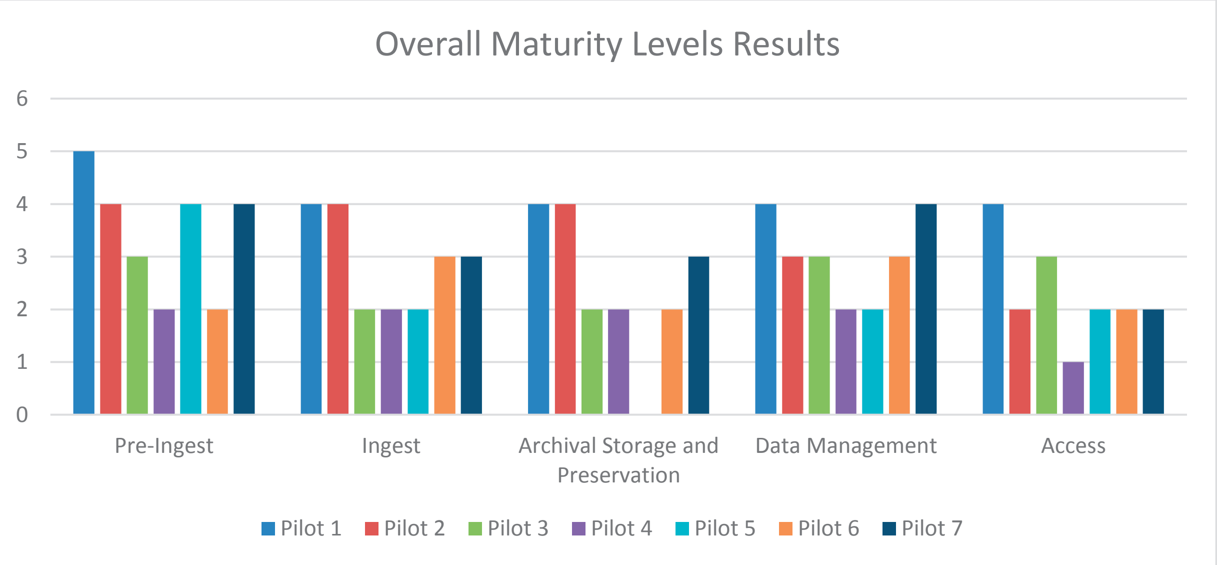


Figure 4. Final Results of the Maturity Levels for All Pilots

Table 2. Final Results of the Answers for All Pilots

Q	Capability / Question Title	P1	P2	P3	P4	P5	P6	P7
Pre-Ingest		**5**	**4**	**3**	**2**	**4**	**2**	**4**
1	Deposit Terms Negotiation	5	5	3	4	5	3	4
2	Producer SIP Validation	5	5	3	2	5	3	4
3	Provenance verification mechanisms	5	5	3	2	5	3	4
4	Enhancement of the Producer SIP	5	1	4	2	3	2	4
Ingest		**4**	**4**	**2**	**2**	**2**	**3**	**3**
5	Creation of fonds	5	1	3	3	-	3	2
6	Ingest SIP verification mechanisms	5	5	3	2	4	5	4
7	Ingest Producer/depositor responses	4	5	4	1	3	3	4
8	Ingest actions and administration processes records	5	5	3	3	5	3	2
9	Legal Rights	5	5	3	1	3	1	3
10	AIP generation procedure	4	4	3	4	1	4	4
11	SIP final disposition documentation	4	5	3	1	3	3	4
12	AIP parsing	1	5	1	1	3	3	4
13	AIP unique identifiers convention	3	5	3	3	3	3	4
14	Preservation Description Information (PDI) acquiring procedures (from a SIP)	5	2	3	2	3	3	4
15	Preservation Description Information (PDI) maintaining procedures	5	5	3	2	3	3	4
16	AIP content information testing procedure	5	2	2	1	-	3	2
17	AIP completeness and correctness	4	5	3	2	3	3	4
18	AIP creation records	4	5	3	3	3	3	4
Archival Storage and Preservation		**4**	**4**	**2**	**2**	**-**	**2**	**3**
19	AIP Storage Procedures	5	5	3	1	-	2	4
20	AIP integrity monitoring	5	5	3	5	-	3	4
21	AIP actions records	5	5	3	2	-	3	4
22	AIP Designated Community Requirements	2	1	1	1	-	1	2
23	Independent mechanism for content integrity checking	4	5	2	2	-	2	4
24	AIP Linking/resolution services	5	2	3	1	-	3	4
25	Tools and resources to provide representation information	5	5	3	2	-	2	4
Data Management		**4**	**3**	**3**	**2**	**2**	**3**	**4**
26	Designated Community information requirements	4	2	3	2	3	3	4
27	Descriptive information association with the AIP	4	5	3	3	3	3	4
28	Bi-directional linkage between the AIP and descriptive information	5	2	3	1	1	3	4
Access		**4**	**2**	**3**	**1**	**2**	**2**	**2**
29	Creation of a DIP	5	1	3	2	3	3	4
30	Access policies	4	1	3	3	3	3	4
31	Access policies compliance	4	1	3	1	2	1	2
32	Access failures and errors	4	5	3	1	1	1	1
33	Access Data Reports	4	-	3	1	3	3	1
34	Access Data Problem/Error Reports	3	1	3	1	2	2	2
35	Access Policies and Procedures	5	5	3	1	3	3	4

Table 4. Overall Results of the Post-Assessment Questionnaire

Aspect	Pilot	Yes	Partially	No
Were the instructions clear and specific?	2	X		
	5	X		
	6	X		
Was the comment box for each question appropriate to complement the answer provided to the question?	2			X
	5	X		
	6			X
Did the assessment cover all the aspects you think that are relevant for Archival Management Practice?	2	X		
	5	X		
	6		X	
Could you relate the aspects being assessed to your pilot context?	2		X	
	5		X	
	6	X		
Did the results of the assessment reflect the current state of affairs in your pilot?	2	X		
	5	X		
	6		X	
Were the assessment results useful as means to check the current state and plan for improvement?	2		X	
	5	X		
	6	X		
Was the assessment a positive experience?	2	X		
	5	X		
	6	X		

This questionnaire was divided into six parts, the first five containing related questions about the different capabilities being assessed. The final part is about overall questionnaire satisfaction. The estimated time require to fill in this questionnaire was 30-40 minutes.

The post-assessment feedback process consists of a set of feedback cycles where in each cycle a limited number of pilots are required to provide feedback. This process allows: (1) to incrementally improve the assessment process, and (2) to manage the pilots' efforts consistently across the last project year. The feedback received from the different pilots was: Pilot 3: Ingest from government agencies (National Archives of Estonia), Pilot 5: Preservation and access to records with geodata (National Archives of Slovenia), and Pilot 6: Seamless integration between a live document management system and a long-term digital archiving and preservation service (KEEP SOLUTIONS).

After analyzing the results of the post-assessment questionnaire the information governance maturity model development team met with the pilots to go over the results of the analysis and address the issues that were detected.

Regarding the overall satisfaction with the assessment, Table 3 details the results of the post-assessment questionnaire questions, related to overall satisfaction with the initial assessment and evaluation. The results are shown for each of the pilots selected to answer the questionnaire.

The results obtained from the analysis of the overall satisfaction with the assessment show that pilots found the assessment a positive experience. However, there are still some aspects to improve, such as the space provided for comments, assessment coverage of information governance and the usefulness of the assessment to plan for improvement. Regarding the comment space, there are already plans to improve this aspect by allowing pilots to include images, and upload documents as a means of providing evidence for the answers given to the questions. Regarding assessment coverage, in the next version of the Information Governance Maturity Model there will new sources of documentation to be analyzed with the purpose of expanding the current coverage of the maturity model. Finally, regarding the improvement plan, we are planning to have the maturity assessment tool provide an improvement plan alongside the maturity assessment results.

7. CONCLUSION AND FUTURE WORK

This paper presented a method to perform a self-assessment of the Information Governance Maturity Model which consists of a toolset consisting of both the maturity model and the self-assessment method which guides the assessment of the state of information governance in organizations as well as provide an improvement path that organizations can follow to enhance their information governance practice.

As future work resulting from this paper, we concluded that current maturity assessment methods focus on highly complex and specialized tasks being performed by competent assessors in an organizational context. These tasks mainly focus on manually collecting evidence to substantiate the maturity level calculation[34]. Because of the complexity of these methods, maturity assessment becomes an expensive and burdensome activity for organizations.

As such, one major area to invest is to develop methods and techniques to automate maturity assessment. Due to the wide spread of modeling practices of business domains, assisted by modeling tools, makes it possible to have access, for processing, to the data created and managed by these tools. Also, the recent state of the art demonstrating how business processes and Enterprise Architecture models in general can be represented as ontologies has raised the potential relevance of the semantic techniques for the automated processing of these models. As such, the objective is to analyze the potential, and the main limitations, of the existing semantic techniques to automate methods for the assessment of Maturity Models through the analysis of an existing model representation of a reality.

There are several examples of models used to represent an organization architecture, such as, Archimate, BPMN or UML. These models are descriptive and can be detailed enough to allow to perform, to some extent, maturity assessment. For example, the collected evidence from an organization can be synthetized into a set of model representations that can then be used when analyzing and calculating the maturity levels.

However, in order for these models to become relevant for maturity assessment there should be a formal representation for both Maturity Models and model representations. One hypothesis is that building on the knowledge of ontologies from the computer science and information science domains, these can be used to represent Maturity Models and model representations. This can be achieved by developing a generic ontology that expresses all these core concepts (or at least a relevant group of them) and relationships among them, as also the rules for a generic maturity assessment accordingly Then, by representing Maturity Models and models representations of concrete organizational scenarios using ontologies we can verify if an organization models representations matches the requirements to reach a certain maturity level using ontology query and reasoning techniques, such as SPARQL and Description Logics inference.

The final objective is thus to identify how these methods and techniques can be used in existing maturity assessment methods, so that they can be proven as relevant to enable the automation of certain aspects of maturity assessment, such as, the maturity level determination. In order to do this, there should be an exploration of what types of analysis can be performed using the information on model representations that is relevant in a maturity assessment effort.

8. ACKNOWLEDGMENTS

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2013, and by the European Commission under the Competitiveness and Innovation Programme 2007-2013, E-ARK – Grant Agreement no. 620998 under the Policy Support Programme. The authors are solely responsible for the content of this paper.

9. REFERENCES

[1] ISO 16363:2012. Space data and information transfer systems – Audit and certification of trustworthy digital repositories. 2012.

[2] ISO 14721:2010. Space data and information transfer systems – Open archival information system – Reference model. 2010.

[3] The Open Group (2011). TOGAF Version 9.1. Van Haren Publishing.

[4] J. Becker, R. Knackstedt, J. Pöppelbuß.,"Developing Maturity Models for IT Management – A Procedure Model and its Application," In Business & Information Systems Engineering, vol.1, issue 3, pp. 212-222. 2009.

[5] A. Brown, Practical Digital Preservation - A how-to guide for organizations of any size, Facet Publishing. 2013.

[6] C. M. Dollar, L. J. Ashley, "Assessing Digital Preservation Capability Using a Maturity Model Process Improvement Approach", Technical Report, February 2013.

[7] A. Tarantino, Governance, Risk and Compliance Handbook. John Wiley & Sons, Inc., 2008.

[8] IT Governance Institute, IT Control Objectives for BASEL II – The Importance of Governance and Risk Management for Compliance. 2006.

[9] T. Lei, A. Ligtvoet, L. Volker, P. Herder, "Evaluating Asset Management Maturity in the Netherlands: A Compact Benchmark of Eight Different Asset Management Organizations," In Proceedings of the 6th World Congress of Engineering Asset Management, 2011.

[10] Real Story Group, DAM Foundation, "The DAM Maturity Model." [Online]. Available: http://dammaturitymodel.org/

[11] ARMA International, "Generally Accepted Recordkeeping Principles - Information Governance Maturity Model." [Online]. Available: http://www.arma.org/principles

[12] ISO 9001:2008: Quality management systems – Requirements. 2008.

[13] CMMI Product Team, CMMI for services, version 1.3. Software Engineering Institute. Carnegie Mellon University, Tech. Rep. CMU/SEI-2010-TR-034, 2010.

[14] A. Tonini, M. Carvalho, M. Spínola, "Contribuição dos modelos de qualidade e maturidade na melhoria dos processos de software," Produção, Vol. 18, No 2, pp. 275-286. 2008.

[15] M. Paulk, B. Curtis, M. Chrissis, C. Weber, "Capability Maturity Model for software," Version 1.1 CMU/SEI-93-TR-24, Pittsburgh, Pennsylvania, USA, Carnegie Melon University. 1993.

[16] E. Anderson, S. Jessen, "Project Maturity in Organizations," International Journal of Project Management Accounting, Vol. 21, pp. 457-461. 2003.

[17] R. Fitterer, P. Rohner, "Towards assessing the networkability of health care providers: a maturity model approach," Information Systems E-business Management, Vol. 8, pp. 309-333. 2010.

[18] A. Sen, K. Ramammurthy, A. Sinha, "A model of data warehousing process maturity," *In IEEE Transactions of Software Engineering.* 2011.

[19] T. Mettler, "A design science research perspective on maturity models in information systems," St. Gallen: Institute of Information Management, Universtiy of St. Gallen. 2009.

[20] A. Amaral, M. Araújo, "The organizational maturity as a conductive field for germinating business sustainability," *In proceedings of Business Sustainability I Conference*, Póvoa do Varzim, Portugal. 2008.

[21] T. Cooke-Davies, A. Arzymanowc, "The maturity of project management in different industries: An investigation into variations between project management models," International Journal of Project Management, Vol. 21, No 6, pp. 471-478. 2003.

[22] D. Hillson, "Maturity - good or bad?," Project Manager Today, March, pp. 14. 2008.

[23] M. Röglinger, J. Pöppelbuß, "What makes a useful maturity model? A framework for general design principles for maturity models and its demonstration in business process management," *In proceedings of the 19th European Conference on Information Systems*, Helsinki, Finland, June. 2011.

[24] N. Brookes, R. Clark, "Using maturity models to improve project management practice," *In proceedings of the POMS 20th Annual Conference*, Florida, USA, 1-4 May. 2009.

[25] C. Demir, I. Kocabaş, "Project management maturity model (PMMM) in educational organizations," Procedia Social and Behavioral Sciences, Vol. 9, pp. 1641-1645. 2010.

[26] OPM3, "Organizational Project Management Maturity Model," Newtown Square, Pennsylvania, USA, Project Management Institute. 2003.

[27] L. Hersey-Miller, "Organizational project management maturity model (OPM3)," Office of Major Projects, Quarter 3, September, pp. 1-8. 2005.

[28] M. Kohlegger, R. Maier, S. Thalmann, "Understanding maturity models: Results of a structured content analysis," *In proceedings of the I-KNOW '09 and I-SEMANTICS '09*, 2-4 September 2009, Graz, Austria. 2009.

[29] A. Korbel, R. Benedict, "Application of the project management maturity model to drive organisational improvement in a state owned corporation," *In proceedings of 2007 AIPM Conference*, Tasmania, Australia, 7-10, October. 2007.

[30] G. Jia, Y. Chen, X. Xue, J. Chen, J. Cao, K. Tang, "Program management organization maturity integrated model for mega construction programs in China," International Journal of Project Management, Vol. 29, pp. 834-845. 2011.

[31] M. Koshgoftar, O. Osman, "Comparison between maturity models," In proceedings of the 2nd IEEE International Conference on Computer Science and Information Technology, Vol. 5, pp. 297-301. 2009.

[32] D. Prado, "Gerenciamento de Programas e Projetos nas Organizações," Minas Gerais: Nova Lima. 2004.

[33] R. Jamaluddin, C. Chin, C. Lee, "Understanding the requirements for project management maturity models: awareness of the ICT industry in Malaysia," *In proceedings of the 2010 IEEE IEEM*, pp. 1573-1577. 2010.

[34] E-ARK Project, "D2.1 - General Pilot Model and Use Case Definition." 2015.

[35] E-ARK Project, "D7.1 - A Maturity Model for Information Governance - Initial Version." 2015.

[36] E-ARK Project, "D7.2 v2 - Information Governance Maturity Model - Initial Assessment and Evaluation" 2015.

Conceptualising Optimal Digital Preservation and Effort

Sean Mosely, Jessica Moran,
Peter McKinney, Jay Gattuso
National Library New Zealand
P O Box 12349
Wellington 6001
+64 4 474 300
{sean.mosely,
jessica.moran,peter.mckinney,
jay.gattuso}@dia.govt.nz

ABSTRACT

In this paper we describe the National Library of New Zealand's attempts to conceptualise how we measure the degrees of effort required to achieve acceptable levels of digital preservation. We argue that understanding digital preservation practice in terms of "optimal effort" may help us conceptualise where and how best to achieve the greatest impact in proportion to effort. The paper examines the various roles of digital preservation, including the archival/curatorial, digital object handling, preservation management, and policy roles through case studies of our experience. We argue that through conceptualising our ideal digital preservation and the levels of effort required to achieve those, we will be able to better understand where our effort is being expended and the levels of preservation we are achieving.

Keywords

Digital preservation; digital preservation roles; effort; ingest; format migration.

1. INTRODUCTION

The mission of digital preservation is relatively straightforward – to ensure that digital objects are kept safe and accessible for as long as they are required. In some sense this mission will always be aspirational. In this paper we will describe some of the challenges inherent in the practice of digital preservation, as well as some of the National Library of New Zealand's (NLNZ) attempts to define a level of comfort in our practice. The discipline of digital preservation demands practitioners be able to acquire digital objects, maintain them in a way that ensures their physical and contextual integrity, and to deliver them for consumption when required. Assuming that the time period for requiring these objects is infinite (or at least undefined), then the task of preservation will never be complete – only once the period of requirement has ended will we know whether our mission was successful for that object. Therefore, our goal is to understand where and how our effort should best be focused.

There is another aspect of our work where we will always be aspiring to an idealised, abstract goal; the relationship of the "original experience" of the object versus how it will be consumed in the future. Regardless of the preservation methodologies employed – migration, emulation, normalisation, hardware/software museum-based, etc. – there may always be some qualitative difference between how the object was previously consumed and how it will be consumed both now and in the future. It is coincidental that of these cited approaches, the one that appears to get the closest to that idealised goal (the hardware/software museum-based approach) also appears to be the hardest to guarantee over time. And even when all technical factors have been controlled, the objects will still have been removed from their original temporal context.

Over time, the actions we perform on digital objects will also threaten them, whether from necessary changes in format, environments, the behaviour of emulators, or some other as-yet unknown factors. Our efforts will always be our best attempt to retain what can never be fully retained—the pursuit of a myth of Total Preservation.

This paper uses the various roles or preservation actors, such as the archival or curatorial, digital object handling, preservation management, and policy, to examine the levels of effort and preservation achieved through two case studies: an ingest of a collection of archival organizational records, and a format migration of a set of WordStar files.

2. TOTAL PRESERVATION?

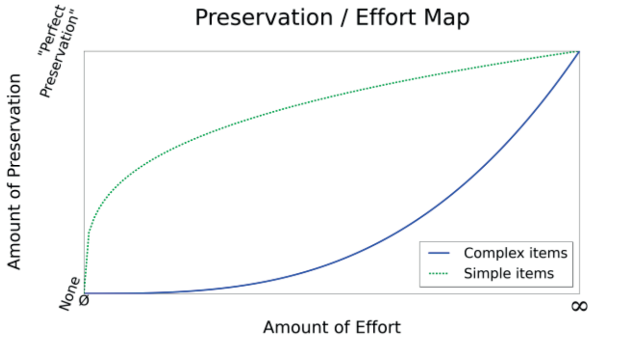


Figure 1. Relationship between preservation and effort over time, for both simple and complex items.

Figure 1 shows a basic illustration of the relationship between the goal of "Total Preservation" and the amount of effort required to achieve this goal. The presumption in digital preservation is that as our efforts increase, be they in the amount of time we take to understand context, identify and validate file formats, ensure we have stable systems to manage and store, and develop and maintain adequate policy to manage our process, the closer we will get to "Total Preservation."

However, it should be noted that in practice the above model does not hold true for all digital objects. For certain types of use cases, it appears that the initial preservation challenges will be very difficult, but that once these initial challenges are ironed out, the effort-to-reward ratio will most likely start to reflect that of simpler objects.

3. OPTIMAL PRESERVATION

3.1 Maximising Outcomes in Proportion to Effort

These two abstractions of effort-versus-reward are intended to reflect two hypothetical use cases – the first being a simple, self-contained resource which was created according to a widely-endorsed standard and can be rendered faithfully in a variety of applications (a tiff image, for example).

The second abstraction is intended to reflect a more complicated use case – such as a multi-file resource that is delivered via a server and involves the dynamic generation of content from a database, augmented by technologies which are being regularly updated or replaced (such as JavaScript-driven web application, retrieved from a LAMP stack and rendered through a modern web browser).

These abstractions, while useful tools for intellectualising the scale of digital preservation workloads, will change depending on the specifics of each use case and the preservation methodologies employed. Later in this paper, when the NLNZ's preservation actions are plotted in regard to this effort-versus-reward spectrum, the exponential curves will be replaced with straight-line steps, in order to situate those actions in a more quantifiable space.

All memory institutions, regardless of their size or the extent of their resources, are affected by the realities of this effort-versus-reward ratio. As the discipline of digital preservation has become more widespread and more institutions begin to address their backlogs of digital content, more practitioners have started to discuss how to maximise their output for their efforts. Such conversations have given way to initiatives like POWRR (Preserving digital Object With Restricted Resources) [10] and the State and University Library of Denmark's Minimal Effort Ingest approach.[5] These initiatives acknowledge the difficulty of adhering to the 'best practice' ideals of the discipline, and the practitioners seek to establish more achievable baselines for digital preservation.

The goal of such initiatives is laudable. By attempting to lower the barrier of entry to the discipline, these initiatives have the potential to encourage additional institutions to implement their own preservation strategies, and start to actively preserve content before it approaches a point of obsolescence. However, the terminology used in such initiatives may be problematic – POWRR's approach of "good enough" digital preservation and the Danish State and University Library's "minimal effort" are couched in language that has the potential to misrepresent the very nature of preservation.

To a degree, this use of diminishing language reflects the broader Information Technology industry as a whole. 'Laziness is a virtue' has long been a mantra of developers and system administrators[1], and the notion of 'good enough' or 'just in time' workflows has driven many large-scale IT businesses.[2] However, whereas digital technologies in general may benefit greatly from an approach that seeks to limit extraneous effort (for example, developing an application in Python rather than C++ when it is determined that the development time savings will outweigh the potential performance gains), preservation is often a different matter. The discipline of digital preservation is still very much in its infancy, and if our language suggests to new practitioners that it is prudent to shy away from the emergent challenges, then there is a much greater risk that the alarmist claims of a 'Digital Dark Age' will become real.

In light of this, the discussions of "good enough" and "minimal effort" should perhaps be reframed as "optimal effort" – in other words, how do we find the best way to measure and maximise our efficiency for preserving digital objects? We want to leave room in our model as well to stress the importance of contributing to investigations into new preservation technologies, as innovations will allow us to preserve more content and further maximise our outcomes in proportion to effort.

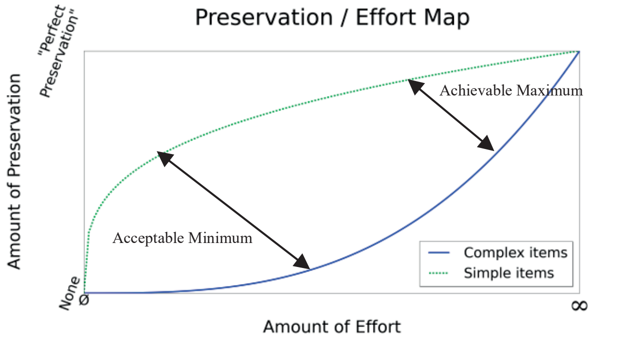


Figure 2. Optimal preservation will tend towards a point between 'acceptable minimum' and 'achievable maximum'.

3.2 Preservation Actors

In order to evaluate our potential for efficiency in preserving digital objects it is helpful to conceptually break up the discipline into its different roles and responsibilities. Even in small institutions where the curator may also be responsible for technical analysis or system administration, the scope for acceptable loss may be different depending on the role. For instance, it may be considered an acceptable risk to undertake sampled quality control of descriptive metadata or access derivatives, but a sampled approach to checksum validation may be unacceptable. For the purposes of this paper, we have chosen to break up the responsibilities of digital preservation into four responsibilities:

- Archival/Curatorial
- Object Handling
- Preservation Management
- Policy

3.3 Archivist /Curator

This role is the traditional agent who advocates for the collection and manages the relationship between the collection, its scope and the contents of individual files and digital objects. This role typically relies heavily on human decision making, using training and experience to understand the intellectual, evidential, heritage, or value of an item. Further, this role should understand the relative impact of any changes, gaps, and other such subjective measures that might be encountered through the processing of any given file.

[1] The three virtues of programming – laziness, impatience and hubris – are widely attributed to Larry Wall, founder of the Perl language. Wall, along with co-authors Tom Christiansen and Jon Orwant, promote these virtues in their highly-successful and influential book *Programming Perl*. [12]

[2] The example of Amazon.com's initial book sales business model (where stock would not be kept on-hand, but rather ordered only once the customer had paid for the purchase) is a canonical example that, for the sake of brevity, will stand in here for a more comprehensive summary.

3.4 Object handling

This role is charged with delivering the technical expectations of the Archivist/Curator role, and ensuring that files are engaged with according to recommendations. The role also provides some technical information and advice to the archivist / curatorial role, helping intellectual decisions to be informed where relevant by technical parameters. We imagine this role to be an even mixture of human decision making, and scripted logic and inference.

3.5 Preservation Management

This role is responsible for the underpinning technologies that bind digital collections together. We imagine this role to manage the Digital Preservation System (DPS) in its widest definition. Parts include digital shelving, per item cataloguing and indexing, processing workflows, and managing other generally automated functions. This role also includes management of digital storage, and regular system updates, testing, and implementation

We imagine this role to be primarily systemic in essence, including the combined processing and feature set of all related applications used to manage and process collections.

3.6 Policy

This role moves across all decision making layers, informing the institutionally agreed processes and functions that are applied to all digital collections.

We imagine this role to represent the collective institutional wisdom and knowledge that determines what can and cannot be undertaken, and the process through which any final outcomes are measured and approved.

3.7 Interpretation

The interpretation continuum (Figure 3) represents the type of reasoning that is required at the various levels to ensure that any interaction or intervention with any given file or digital object is being properly handled. The Archival/Curatorial role is predominantly interested in the intellectual content, context, provenance, and chain of custody of objects. These concerns include: what the digital object represents, its expected access by readers, its relationship to other objects in the collection, and how pre-ingest and ingest activities may affect an object's authenticity. The Object Handling role provides information to the Archival/Curatorial role on the technological possibilities and limits. This role also works closely with the Archivist/Curator to ensure digital objects are handled properly and technical solutions are developed. The System role is predominantly concerned with the technical representation of the object – what encodings are being used to bind information to any representation, what processes or operations are permitted and how they are carried out, how the host operating system and file store understands the binary stream and its attendant metadata. The Policy role helps develop the principles and directions to which the other roles will work.

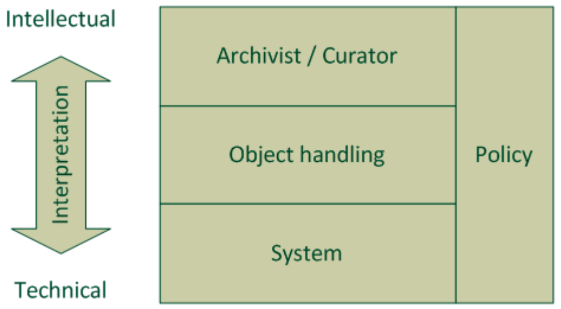


Figure 3. Interpretation continuum.

Figure 3 attempts to take the notion that for any unit of effort, many files may be processed in a light or basic way, or few files may be processed in an intensive or complex way, and understand the way that all the roles rely upon and work together. In this construct, effort is seen as a mixture of resources (people, time, money) and capability (skill, tools, knowledge).

Essentially, while we conceptulise these roles as separate and independent, in effect they must work together, bringing their different expertise's to bear on the decision making processes. As we understand it, our work is continually being informed by both policy and the intellectual and technical parameters necessary to achieve what we think of as optimal digital preservation at any one moment.

4. APPLYING CONTINUUM TO A COLLECTION

4.1 Collection Description

To understand how we are applying this continuum model in more detail, it is helpful to apply it to a sample collection from the Library. This sample collection consists of the business records of an organization and was transferred to the NLNZ in 2013. The records came to the Library in two transfers over the course of two months with the organisation's IT department transferring the records from shared drive storage to external hard drives supplied by the Library. Prior to the transfer, a Curator and Digital Archivist visited the organisation, interviewed the donor about the kind of materials to be transferred and were given a high-level overview of the records. Together, we selected the areas of the shared drive the Curator appraised to have transferred to the Library based on the Library's collection and digital preservation policies.[7] Like other institutions, we believe time and effort should first be expended to develop policies around what and how we will collect and preserve digital content. We rely on these policies to guide our decision making throughout the appraisal and ingest workflow.[1] Based on this visit we suspected the records to consist largely of business records created in a Windows environment, using standard and well-supported file formats including the Microsoft Office Suite, pdf, tif, and jpg files, and most created in the last 10-15 years. At this stage our understanding of the collection was based only on an initial visual appraisal of the records.

4.2 Technical Appraisal

Once the collection was ready to be transferred to the Library, technical and intellectual analysis of the collection began. We determined that the collection consisted of 4239 individual files, and while at the top levels the records were well organized, in many cases the file paths were five or six levels deep; the collection had a total of 355 folders and a total size of only 4 GB. The collection dates ranged from 1997-2012.

We expected that because these were current business records, maintained by the organisation, and using widely adopted file formats, that our digital preservation challenges would be minimal. However, during technical appraisal the digital archivist discovered that the collection included 316 files with mismatched file extensions and 10 files whose format was unidentifiable,[3] as well as a number of files with special characters in the filenames.

[3] The Library uses DROID for format identification as part of its pre-ingest technical appraisal and within its DPS.

The digital archivist at this pre-ingest phase had a number of decisions to make in terms of how to best prepare the collection for ingest into the digital preservation system (DPS). The collection had to be appraised, and arrangement and description performed. These processes were done by subject experts in those roles with advice from the digital archivist. In this case, because the collection had been transferred intact from the shared drive of the organisation, there were some records that the Curator did not want to collect and preserve. Based on a desire to retain the deep file structure of the collection the digital archivist worked with the arrangement and description librarian to describe and further appraise the collection. At these points the digital archivists provide technical advice to the curator and arrangement and description librarian about the types of file formats in the collection, how they were created, and how best to described them for future preservation and access. In this work the digital archivist is informed by the Library's digital preservation policy as well as both intellectual and technical knowledge about the records and an understanding of the DPS system and its strengths and limitations.

4.3 Preparing for Ingest

A number of policy and business decisions inform how we ingest material into our DPS. These include: records going into the DPS will have at least a scant collection management record, and files should pass validation checks, including format identification and validation. While neither of these is necessary for ingest, the Library has made the policy decision that by doing this work at ingest we are better prepared to understand our growing collections overtime and better able to make decisions about what sort of preservation actions we may need to perform in the future.[9]

Once appraisal and arrangement and description were complete, we were ready to being the process of ingesting the collection. First, we identified those files which could be easily ingested via the Library's internal submission tool already developed and integrated with the DPS. Using this tool, the digital archivist is able to build Submission Information Packages (SIPs) that automatically deposit the files and metadata to our DPS. In this case we selected those files which would need no preconditioning or provenance notes added to their preservation metadata, and that were all part of the same folder groupings. If these two conditions were met they could be quickly deposited using our ingest tool. This method accounted for 966 IEs or 23% of the collection, and in this case the greatest effort was expended earlier in the development of the ingest tool. Next, the digital archivist filtered out from the remaining records all those files that had been appraised out of the collection during appraisal or arrangement and description. This accounted for another 705 files, or 17% of the collection. However, that left us with about 60% of the collection that could not be quickly or easily ingested. In this case, that was due to the organizational structure of the files, lack of format identification, mismatched file extensions, or some combination of the above. At this point the digital archivists handed the collection over to the preservation analyst to do more of the object handling and determine the best way to ingest the rest of the collection. The 60% of the collection represented about 2500 files, and while this is still a relatively small number of files, we deemed it too large a number to be ingest using our ingest application, because in order to retain both the file structures and apply the preconditioning and provenance notes would mean hand building hundreds of individual SIPs. We deemed this to be too much manual effort. Instead, we developed a script that could identify some of the main issues that would cause the files to fail validation during ingest, automatically address those issue that could be fixed, apply the accompanying provenance notes, and prepare the files for ingest.[6]

This second round of ingest accounted for another 2429 IEs, or 57% of the collection. In this part of the ingest process most of the time and effort was taken in developing and testing the script and data wrangling to prepare the files for ingest. We now had less than 200 files remaining that could not be ingested. Some of these were more complex multi-file IEs identified during processing and loaded separately by the digital archivists. The remaining files included 4 files that can be loaded following the next PRONOM update, 12 files whose format we have been able to identify, but do not yet have signatures written, and 5 files whose format we could not identify and that still require further research.

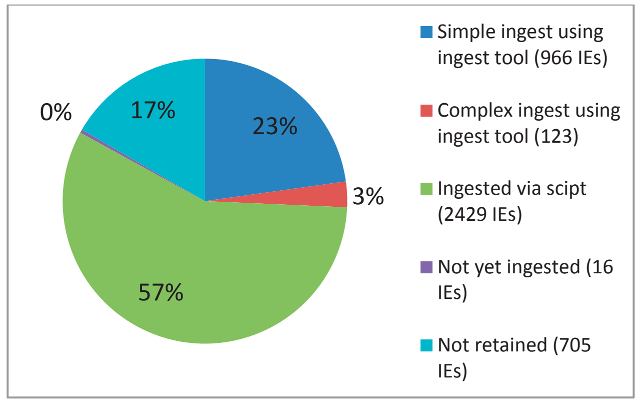


Figure 4. Types of ingest in collection.

Figure 4 illustrates the percentage of the collection by ingest type, and hints at the effort expended by type of ingest. For the first batch of 966 IEs (which were deposited by the simple ingest method), the effort in that case came in the development of our ingest tool and its stable integration with our system. In other words, this ingest method was simple because all the tools were already in place. Next, the 2429 IEs ingested via a script required more upfront effort in understanding the objects, developing, and testing the script and the automated ingest method. Once that development and testing effort has been expended we anticipate being able to transfer the knowledge and tools developed in this collection for use in other similarly complex collections. The remaining 3% of the collection required the most effort, through manually preparing the files for ingest, format identification, writing of signatures, and other object handling.

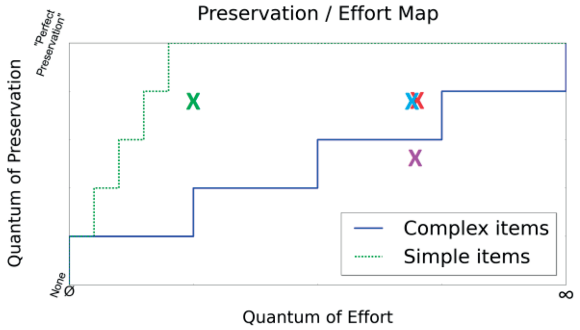


Figure 5. Effort mapped for sample collection ingest.

Figure 5 shows how we see the various parts of the collection mapped against our continuum of effort and preservation. For those parts of the collection that we were able to ingest with relatively little effort using tools already in place, we think we were able to reach an acceptable level of preservation with less effort. For those files where we had to either build more complicated SIPs using our ingest tool, or develop a scripted process to prepare the files for ingest, a much higher degree of effort was required to reach a similar level of preservation. The remaining files (those which have not yet been ingested into our DPS) have required more effort, but a much lower level of preservation has as yet been achieved. We can extend the same thinking across collections (Figure 6). For many of the collections that come into the Library and would not be categorised as complex by us, with less effort we can be confident in a high degree of preservation. For those collections that take longer to appraise and process, either due to size, file format, or condition of the collection, greater effort is needed to reach the same level of confidence in our level of preservation. Indeed in some cases we have already expended a great deal of effort for a much lower degree of degree of preservation.

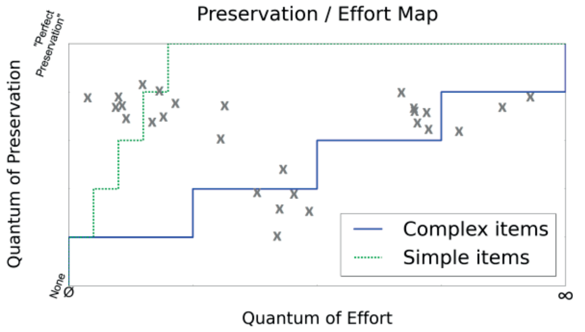


Figure 6. Effort mapped for ingest across many collections.

4.4 Preservation Management

Once the collection is ingested into the system, it then comes under the purview of preservation management. The activities in this area are in place to ensure that the objects remain free from corruption, are available and can be accessed over time (while maintaining any access restrictions).

Some of this work can be system automated. This includes routines such as periodic virus scanning and checking of fixity values. Other processes, however, are not automated. These include risk analysis and preservation planning and actions. Automation is not available for these processes due to the immature nature of the work. For example, we have recently undertaken a migration of WordStar files [4]. This migration was a handcrafted solution, both in terms of generating tools from scratch and taking deliberate time and care with decision-making. It was handcrafted because: a) existing tools for the conversion were untrustworthy; b) we wanted to ensure the process was robust; and c) non-technical staff required time to understand the activities we were undertaking and had to be assured that the resulting files could stand as authentic representations of the originals (and conceptually replace them).

NLNZ is embarking on a programme of migrations. One of the outcomes of this programme (beyond the primary concern of mitigating risks) is that a clear process, including decision-making routines, is agreed upon. With this in place, far less effort will be required to achieve the same outcomes as the WordStar conversion.[4]

In addition to these management aspects, there is of course the underlying architecture upon which the management takes place. A preservation system is not (at least in our experience) a "plug-and-play" operation that can be left to its own devices. The underlying storage is reconfigured regularly, as is the processing layer. Such reconfiguration is due most often to the availability of new technologies, changes in requirements, new external vendors, and collection growth. There are also updates to the preservation system itself. Requirements need to be written, negotiation over the solution undertaken, testing, and eventual roll-out.

Preservation management can be best described as a container for iterative processes undertaken across the lifetime of the objects being cared for – the "lifetime", in our case, being perpetuity.

4.5 Access

Access is in fact an added benefit of our current process and one not always discussed in digital preservation literature.[8] One of the benefits of expending this effort upfront to identify, prepare, and validate all the files that go into our DPS is that they are accessible immediately to researchers. Because the files have been identified, validated, and have the correct extensions, files in our DPS system are accessible and can be delivered to our users either through viewers or they are downloadable and accessible to users, even those with little technological confidence. By performing our quality assurance and pre-conditioning actions during ingest, the files can be delivered back to our researchers in a format that their computers are more likely to understand with little or no intervention from us.

4.6 Policy

Underpinning all of this work is the policy layer. Policy informs and aids decision making at each step of the process. Our policy is aimed at the level of operating principles[5]. That is to say, we describe the goals and aims for policy areas and describe the high level processes that should be undertaken and within what boundaries. However, the policies do not go into highly detailed specifics. For example, the fixity policy will contain principles detailing that more than one fixity information method must be used, but it does not specify which ones should be used.

Each policy created requires a large amount of effort to create. There is consultation, drafting, further consultation, redrafting and, if lucky, sign-off (our experience tells us that this process has to be tightly controlled in order to avoid a constant spinning in the drafting and consultation phase). After a few years, there is also the review of the policies. No little effort is expended across all preservation stakeholders while creating and reviewing these policies.

While policies bring us towards optimum effort through normalizing and codifying practice, they do not completely diminish effort. The policies do not define the exact steps that must be taken by each staff member; they are not business process models. Therefore there is still some layer of effort by the member of staff as they put into practice policy principles.

[4] Indeed, we are even testing a migration that could be classed as "quick and dirty". This should help us explore the boundaries of our comfort around the integrity and authenticity of the content and what measures are required to give this comfort.

[5] For an excellent discussion on preservation policies and the different levels they operate at, see Sierman, 2014.[11]

5. EFFORT

The 'effort' axis on all of the charts above is unencumbered by any sort of scale. Effort can be measured across many different parameters and given many different scales. It could be staff hours, costs involved, or even perceived effort (a la the Borg rating of perceived exertion [3]).

In our conceptual world, effort is predominantly envisaged as staff hours (and perceived staff hours). But it also contains a trace of costs (some of our tool development is outsourced and therefore monetized in a way that our team efforts are not). Hidden in there are also costs of storage and consultancy, which we interpret as part of the effort of preserving.

It seems pertinent to note that even if the scale measured defined costs it is very hard to get a true sense of what, for example, staff resources are being spent on. If an organisation employs two members of staff and assigns them key tasks it is, we would argue from experience, difficult to actually gauge the time spent on, say, format identification, or risk analysis, or the validation of preservation actions. Additionally, are these areas of work more valuable than the support work that is spent wrestling with recalcitrant tools, or tracking down bugs in code, re-architecting server configurations, and testing system upgrades? This is one of the reasons we are reluctant to lay down any definite scale for effort.

This paper is deliberately vague about effort and therefore also about costs. These should be understood in relation to the reader's own organizational context. The determination of exact costs for digital preservation activities, and thus, defining an exact scale on the "quantum of effort" axis would require an in-depth community-agreed process for digital preservation. Even delving into small aspect of the preservation process highlights differences in practice that would change costs. For example, some institutions require and exact single format identification for each object, thus requiring in some cases relatively extensive research to be undertaken. Others accept all possibly format identifications including "format unknown". The Blue Ribbon Task Force have worded this far more eloquently suggesting that arriving at an estimate for preserving an amount of data over time is "a task over-laden with the details of a particular implementation and preservation context" [2].

Likewise, the notion of deferred costs has not been explicitly addressed. This is not to say that there are not ongoing discussions about the upfront costs and delayed costs. Should we be spending *n* amount of days at the point of ingest on certain issues, or rather bring in the material as is and resolve any issues at a later date? Our experience tells us that that when we leave something for a later date, that later date rarely occurs. Our current (non-written) policy is that we must give best efforts now in order to give future staff (and even our future selves) the best possible chance of continuing to offer access to the material.

The final word to address is "efficiency". Efficiency has very good associations to higher rates of productivity, but also has quite negative connotations of adequacy and trimming of outcomes. For the sake of this paper, we will focus only on the positive and consider one final example.

We have mentioned above the work undertaken on converting WordStar files. A great deal of effort was expended in that process. It is clear though that we could have made that process far more efficient: we deliberately slowed much of the process down in order to guarantee that all stakeholders followed every single step of the process even if they were not directly involved in each step. The final graph below shows how we could have put far less effort in for probably exactly the same outcomes (in terms of the content). But what the graph (figure 7) does not show us is the institutional outcomes that were achieved by taking this slowly, slowly approach.

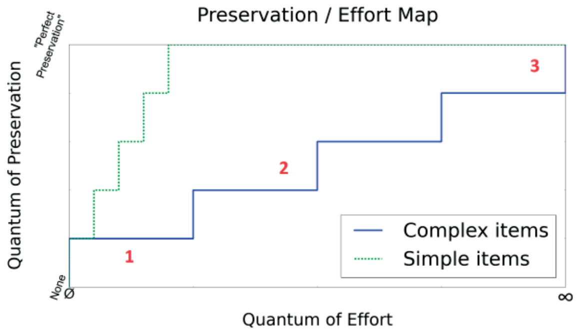


Figure 7. Effort mapped against a single format.

In the graph above we have set the levels of preservation as a very coarse series of descriptors; No preservation; Binary only; Format Identified; Generally Renderable; Obsolesce risk migrated. This is a rough scale for demonstration purposes, and not a measurement we would specifically advocate for.

The quantum of effort is also an approximation, based on the work required to successfully ingest and migrate the original content.

The first marker (1) indicates the state we found the collection in. We could have ingested the original WordStar files as unknown binary items with minimal effort, but as this is against our general working methodology and business rules of only ingesting content that has a format identification we can use the map to indicate where we start.

The second marker (2) indicates the effort required to create a successful format signature, get that included into the format registry, and ingest the content into the preservation system.

The third marker (3) indicates the effort required to migrate the content from the original format into its new contemporary format.

6. CONCLUSION

By conceptualising where our idealised digital preservation is, and what levels of effort are required to achieve it, we can better understand where we are currently expending our effort and what level of preservation we are achieving. Charting this effort and preservation will allow us to begin quantifying what we are doing, what direction we want to move in, and how best to expend our effort to achieve better efficiencies in our digital preservation work. What we do believe (but won't be able to test until some future time) is that the effort we expend will result in the National Library of New Zealand being able to deliver a digital collection back to a user in a way that they can understand its organisation, its context, have trust in it its authenticity, and can easily access the objects and their metadata into perpetuity.

7. ACKNOWLEDGMENTS

Thank you to other staff in the Library whose input has shaped our thinking including Leigh Rosin, Gillian Lee, Susanna Joe, Sholto Duncan, Kirsty Smith, Cynthia Wu, and Suman Sajwan.

Mapping Significance of Video Games in OAIS

Rhiannon S. Bettivia
School of Information Science
University of Illinois at Urbana-Champaign
rbettivi@illinois.edu

ABSTRACT

In this paper, I explore the concept of *significant properties* and how such properties do and do not fit within the Open Archival Information System (OAIS) Reference Model. Combining interview data from research about the deployment of OAIS in cultural heritage institutions with data about video game production and preservation from the Preserving Virtual Worlds II (PVWII) grant project, this paper maps stakeholder-identified significant properties onto the 2012 version of OAIS [4]. Significant properties have many definitions and even many names. Operationalizing this term broadly, many such properties do fit within existing OAIS entities. However, significant properties that are relational and external to digital objects' code and environments do not. This paper concludes that additional metrics are needed to begin shaping the process of documenting significant properties at scale.

Keywords

Significant Properties; OAIS; Digital Preservation.

1. INTRODUCTION

I explore the concept of *significant properties* and how these do and do not fit within the Open Archival Information System (OAIS) reference model. *Significant properties* is not an OAIS-specific term. Operationalizing this term represents a point of tension between the various disciplines brought together to construct the sub-discipline and profession of digital preservation. Significant properties are important because they refer to some kind of information without which digital artifacts are unintelligible, even if the artifacts remain functional. Significance is determined by a variety of stakeholders [7], and digital repositories will not be able to engender the trust of users if they cannot communicate back those elements about a digital object that consumers find most important. Addressing significant properties poses a challenge because they are undefined, or perhaps over-defined: even the term itself is under dispute and there are a variety of alternatives that implicate significance, from Information Properties to significant characteristics to context information. This paper adds to the on-going conversation by identifying significance of digital objects as defined by practitioners and content producers. By beginning with what is described by these stakeholders as essential in particular case studies, this paper attempts to operationalize the concept of significance rather than weighing in on the various definitions. In this way, this research is productive of new possibilities: it is not simply a criticism of existing information structures, but aims away from silo-ing this discussion according to institution or discipline type towards more macro understandings that can inform the creation of metrics to guide processes of documenting significance at scale.

I describe digital preservation as meta sub-discipline of the meta-discipline of information science [1]: in the same way that information science is imbricated across the traditional research disciplinary spaces of humanities, social sciences, and natural sciences/mathematics, thus incorporating and informing all of these areas, so too is digital preservation a meta-sub-discipline of information science: preservation work is part and parcel of all the work information professionals do, and so borrows terms and practices from all areas of the information professions. Significant properties stem from library and archival traditions, yet need to be rendered functional in the broader space of digital preservation, and this poses a challenge that is expressed by the myriad definitions, readings, and projects that reject significant properties as unimportant or untenable. Webb, Pearson, and Koerbin [32] of Australia's National Library sum up this ethos within the general realm of digital preservation:

> *"We have come to a tentative conclusion that recognising and taking action to maintain significant properties will be critical, but that the concept can be more of a stumbling block than a starting block, at least in the context of our own institution."*

This simultaneous acknowledgement of the critical yet poorly understood nature of significant properties demonstrates both the importance of the term, but also the barriers to its productive impact given a lack of definitional clarity: significant properties have become an elephant in the room for digital preservation. I argue that one method of synthesizing these various definitions is to engage with how this term is used in practice.

This paper marries data from semi-structured interviews about the deployment of OAIS within memory institutions with interview data collected during the Preserving Virtual Worlds II (PVWII) grant project. The OAIS interviews cover a range of digial preservation scholars, practitioners, and OAIS authors, revealing insight into how 'insiders' perceive the significance of digital objects. The PVWII data explicitly examine significance as described by content producers, in this case programmers and others working on the creation of digital games and virtual worlds. I examine these data alongside the Transformational Information Properties proposed as an alternative to significant properties in the 2012 version of OAIS [4][13][30], to see how well user-described significance fits within the entities for Representation Information, Provenance, and within the OAIS conception of authenticity. This paper examines how complicated multi-part works like video games, virtual worlds, and other dynamic popular culture materials fit within OAIS. I work with OAIS given its ubiquity in the field. I argue through these data that some significant properties fit within the entities of the OAIS reference model, particularly those related the digital object itself and the software/hardware environments required to make an object functional. However, I also argue that OAIS, as currently scripted, cannot encapsulate all the types of significant properties derived from the interview data. The places where these mismatches occur are places wherein other preservation practitioners and scholars have identified weaknesses in the model related to the changing landscape of digital content towards more distributed models. By deriving importance and productive definitions of significant properties from practitioners, I locate significance in relation to the digital object and identify the types of significance not currently covered by prominent models and advocate for new guidelines that incorporate these.

2. LITERATURE REVIEW

Webb, Pearson, and Koerbin [32] sum up the consensus that significant properties are important, yet difficult to employ for preservation purposes: their description of significant properties as a 'stumbling block' indicates that previous attempts to clarify this term and provide methods by which to make it operational have not been widely adopted. The lack of a simple and widely accepted definition is one difficulty in actually evolving the term *significant properties* into concrete preservation and curation strategies. General discourse on the topic refers to properties that are most essential to the understandability of digital objects over time. That is to say, significant properties recognize both the situatedness of digital artefacts and the fact that it may not be possible or practical to save every aspect of every object over time.

The term *significant properties* has been used in digital preservation and curation literature for over a decade. The most commonly referenced definition, and also an early one compared to others I reference here, is the one by Hedstrom and Lee [14], who define the term as "those properties of digital objects that affect their quality, usability, rendering, and behaviour". *Significant Properties* are described variously in many places, and Giaretta et al [13] and Dappert and Farquhar [7] discuss the difficulty in settling on a single definition by exploring some of the myriad definitions that currently exist in disciplinary literature. These definitions stem from various sources, across institutions, information types, and research disciplines. Of science data, for example, Sacchi et al [26] say:

> *"Although this notion has clearly demonstrated its usefulness in cultural heritage domains, its application to the preservation of scientific datasets is not as well developed."*

What precisely is meant by "demonstrated usefulness" is not entirely clear, as many practitioners in cultural heritage acknowledge the use-value of this notion without being able to advance either a concrete definition of what it means or how to account for it formulaically or machine-readably.

Within the interview data that I present in this paper, definitions of significant properties were similarly varied. One participant from my OAIS interviews, a manager of digital preservation at a European national library, suggested that libraries are well equipped to deal with significant properties, "because…as a library we have a lot of experience in describing things so we are very good at metadata". This quote suggests that she perceives a relationship between descriptive metadata and significant properties. Demonstrating the salience of findings about the occasionally contradictory nature of various definitions of significant properties, another OAIS interview participant, a research and development officer at a European national archive, said "well, [the term *significant properties* refers to] just technical metadata, [doesn't] it?"

The other difficulty with this term is that it represents a larger schism within the field of digital preservation between practitioners from computer science and those who come from archival or library science. Bradley [2] presciently said:

> *"'All God's children got significant properties,' we can sing in unison, but this takes us no further if we cannot define its meaning in such a way that we understand what properties are under consideration, and describe them in a way that is machine-readable and automatically actionable."*

This encapsulates the tension between the social, the human and the technical. Because all of these elements are at play in preservation, particularly when it comes to the preservation of cultural heritage and popular culture materials, significant properties serve as a potential flash point within larger preservation discourses that arise around OAIS and the growth of the field of digital preservation.

The OAIS reference model has long and wide adoption within the digital preservation community. Further, the terms contained therein have come to function as boundary objects across different types of preservation and curation endeavors [22]. Giaretta et al [13] examined the relationship of significant properties to existing entities in preparation for the 2012 revisions to OAIS. The authors proposed a number of existing, and thus more precisely or homogenously defined, terms from within the OAIS reference model to act as an alternative to proposing a new definition for significant properties or reconciling existing ones. They also proposing the Information Property as an alternative. The Information Property and the resultant emphasis on authenticity relies heavily on the *Designated Community* term within OAIS, as authenticity does not exist in a vacuum but is instead a product of the relationship between a potential end-user and the data they might receive from an OAIS. This echoes work by scholars like Dappert and Farquhar [7] who posit that significance is not inherent to objects but determined by stakeholders. The term *Designated Community* is 'weakly defined' in OAIS, according to an interview subject, in the sense that the model does not concretely detail how to form and document such a community. While such specificity is not necessarily within the purview of a reference model, the missing piece with digital preservation practice is that standards subsequent to the reference model have not yet been developed, and that many institutions have not, at a site-specific level, formally defined their Designated Communities [17] [3].

Work that does not address the Designated Community cannot address the significant properties elephant. In dealing with a concept like significance, it becomes necessary to ask *significance for whom*, something that is often implied but not always specifically addressed in discussions of significant properties. Yeo [33] sums this up eloquently:

> *"However, the determination of 'significant properties' is no less problematical than the debate about notions of value …not least because different user communities will bring different perceptions of what constitutes significance."*

The situated nature of the Designated Community and the idea of 'preservation for someone' arise from the same discourses of place and time that inform conversations about significant properties. Struggles I identify here are due in part to changes in technological landscape the importance of which authors of OAIS were not able to predict. This is not new: for example, earlier versions of OAIS assumed migration to be the default preservation method, yet recent years have seen a shift away from migration and normalization towards a more mainstream acceptance of emulation and the importance of computing environments, particularly in reference to complex media like video games [15][23][9][8][6]. The 2012 OAIS revisions encapsulated this change. Recent developments in areas like linked data and other forms of distributed content pose a challenge to the current iteration of the OAIS reference model, and practitioners like David Rosenthal [25] have made calls for attention to this as OAIS heads into a new round of revisions in 2017.

The 2012 changes to OAIS resulted importantly in the definition of the Transformational Information Property, which does some work to capture significance in relation to stakeholders [4]. Sierman [30] compares the most recent version of OAIS with its predecessors and notes:

> *"The Information Property is related to the commonly known but not always clearly defined term "significant property", but I think more discussion is needed to define better where the differences and similarities between the two concepts lie and how to translate this into the daily practice."*

The Transformational Information Property in the 2012 revisions of OAIS is meant to stand instead of significant properties, rather than in place. During my interviews with OAIS authors, some noted that they decided to side-step the *significant properties* discussion entirely by creating a separate entity that would serve a distinct set of functions partly because of the sheer number of incommensurate existing definitions for significant properties. The key is that Transformational Information Properties are meant to work in conjunction with other existing features in OAIS, in lieu of actually defining significant properties, thus avoiding the need for authors and OAIS as a sociotechnical network to engage within this space. In practice, the outcomes are not so neat: by choosing not to wade into the significant properties debate, the OAIS authors are taking an effective stand indicating that the concept does not need to be incorporated within the major standard of the field: significant properties are not useful or important enough. This dictates in part how well significant properties can be taken up by others given the pervasiveness of OAIS and the ways in which practitioners in interviews struggled to envision alternative frameworks for their preservation work. Additionally, the solution conceived of by the authors to avoid the term has not stopped practitioners within the profession from continuing to call for OAIS to deal with significant properties more explicitly.

The treatment of significant properties within the literature is reflective of current discourses in digital preservation practice. As such, suggested models or practices fit squarely within existing models like OAIS and address property/value pairings in relation to aspects of digital objects that are better understood within the general field of digital preservation. This means that earlier literature focused on aspects of digital objects like semantic representation and functional bits; more current work incorporates the environment of the object as well. What this reveals is an additional difficulty in developing means of documenting significance at scale: theoretical approaches focus on significance of the features of digital objects that are prominent in the moment. If, as I will argue here, significant properties of digital objects are located elsewhere, then the current theoretical approaches will not be able to sufficiently account for significance.

3. METHODS

3.1 Data Collection

This paper utilizes two sets of data that capture different views on significance from important stakeholders in digital preservation. The first is comprised of semi-structured interviews conducted with a variety of preservation practitioners, scholars, and OAIS authors. These interviews were conducted in Europe and North America as part of a research project that investigated the effects of OAIS on values and professional practice in cultural heritage institutions. Interviewees included 28 participants from 5 countries. These participants included digital preservation specialists who practice or research in public and private universities; public and national libraries; national and private archives; museums; and consulting firms. Also included were authors of OAIS and data curation scholars working predominantly in the sciences. Within the practitioner interviews, participants had a range of specialties and areas of expertise, including technology officers, research and development administrators, as well as some analog archivists and librarians who had little to no knowledge of OAIS despite working within institutions or departments that are heavily influenced by OAIS. In conjunction with the interview data, this dataset includes a variety of documents such as the various versions of OAIS itself as well as a number of procedural and policy documents submitted to me by interview participants. These interviews were qualitatively coded for characterizations of OAIS; discussions of particular OAIS terms; and descriptions of what is well-enabled by OAIS as well as what is missing or constrained.

The second dataset was gathered as part of the Preserving Virtual Worlds II grant. PVWII was funded by the Institute of Museum and Library Services (IMLS) and concluded in 2013. It included investigators from the University of Illinois, the University of Maryland, Rochester Institute of Technology, and Stanford University. Investigators examined the concept of *significant properties* as it applies to video games with the aim of informing preservation practices for complex media, building on previous projects that examined the significant properties of software and a previous game preservation project, Preserving Virtual Worlds I (PVWI) [19][16][21]. Broken into two investigative phases, Phase 1 entailed a two-fold method for examining significance. Investigators performed technical and content analyses of a set of video game series. Simultaneously, investigators conducted interviews with people involved in the design and dissemination of games from the case set; with designers working in other game design studios; and with fans and programmers who have worked on more well-known modifications (mods) of some of the games from the case set. These interviews were qualitatively coded and analyzed by members of the research team across the various institutions involved in the grant project. Phase 2 of PVWII focused on the development of tools and metrics to assist in the preservation of the significant properties identified from the research in Phase 1. These included an examination of how such properties could inform decisions about the emulation, migration, and re-implementation of games as well as defining benchmarks for authenticity in playback. PVWII suggested a layered model for looking at games, delineating different aspects of each system wherein different users might locate significance. I will discuss this model in greater detail later in this paper.

For this paper, I coded both datasets using NVivo software. There were three overarching nodes: explicit mentions of significance; implicit mentions of significance where participants mentioned terms identical or similar to those that appear in the myriad definitions of the term; and things that were explicitly defined by participants as not significant. Within these first two nodes, responses were further categorized according where significance was located within the layered model mentioned above as well as within existing OAIS entities according to Giaretta et al [13]. The last node recognizes that an equally important part of creating adequate preservation information packages is determining what information should not be saved, and this echoes on-going discussions in the realm of science data curation and media art preservation.

3.2 Research Questions and Process

The research question for this paper is: given the ubiquity of OAIS, how do complicated multi-part works like video games, virtual worlds, and other dynamic popular culture materials fit within the model? I investigate this by allocating significant properties to existing OAIS entities and identifying those which do not fit within the model. I began the process with the hypothesis that all would fit despite the fact that video games and other complex digital objects pose a challenge to digital preservationists for two reasons. First, as mentioned above, the large and general category of significant properties is one that OAIS intentionally avoids. Second, while the term *Archive* in OAIS is very specific, it shares some foundational tenets with the study of traditional paper archiving practice and it is not the job of the traditional archive to collect or preserve external significant properties, those not contained within code or computing environment. Several interview participants expressed views about the traditional role of archives that indicated archives should not collect external significant properties. Within traditional archival practice, the term *selection* is used very narrowly: the scope and collection statement determine the type of content to be collected, and all such content from the organization is archived rather than an archivist selecting certain materials for processing and preservation in a more colloquial sense [10][11]. Creating Information Packages for complex media requires some measure of this latter selection: the Archive must choose a set of things to include in the package that encompasses the most significant properties. The purposeful and transparent creation of artificial boundaries is at odds with foundations of archival practice which inform the authors of OAIS and how OAIS gets deployed. This is the second difficulty that arises when mapping video games into OAIS.

I focus on interview data related to two games franchises from the PVWII case set: Carmen Sandiego and Civilization. For both games, multiple creators were interviewed, painting a broad and varied picture of significance as determined by creators. In order to determine how well this data can be captured by the high-level entities detailed in the OAIS documentation, I parse the data to look specifically for information that could be modeled as Representation Information, especially for the documentation of Provenance; to act as benchmarks for authenticity; and what can be modeled as Transformational Information Properties.

Finally, I identify significant properties that do not easily fit within the Representation Information of particular digital objects and discuss why it is that these do not work within the current iteration of OAIS. Some of these properties are related to the tricky OAIS term Knowledge Base; others are distributed in a manner that challenges OAIS's requirement for adequate control of the content.

4. FINDINGS

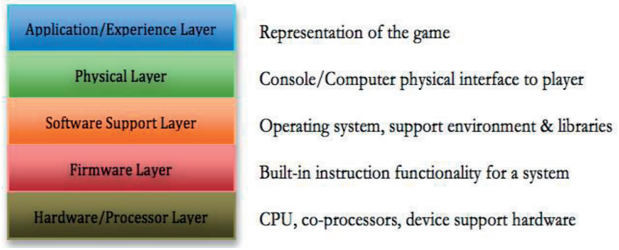
The current interest in emulation as a preservation method does in some measure move the preservation community towards an acceptance that things beyond the object themselves are significant and require preservation. In the case of emulation, significance is found in the behaviors of the original computing environment and this has been recognized in a number of research endeavors including some that specifically examine significance [7][9][6][12]. PVWII research painted a very complex picture of significance within the realm of games. A key finding was, unsurprisingly, that significance is highly situated. The research data indicated that what is significant about games may not be something inherent to the game's code (bits) or even computing environment (platform, operating systems, controls), but could include elements as varied as underlying data models or general surface affective experiences. I argue for the consideration of even broader data about significance that may encompass social and cultural aspects and elements of the Designated Community's Knowledge Base. These terms within OAIS acknowledge that there is more involved in understanding objects than simply recreating the objects themselves: artefacts are a product of a particular place and time, and are understandable as such. Singling an object out as divorced from is spatial and temporal context will not guarantee the understandability of the object over time, even if its rendering environment and bits are preserved: a digital library director at a private US university summed it up nicely saying, "I mean, files are not that useful without something." That is to say, we need something beyond even working files themselves.

The situation (that determines the situatedness) of an object needs to be preserved. I argue that this is what is encapsulated in OAIS by the terms *Knowledge Base* of the *Designated Community* within the sociotechnical complex of OAIS, even if the explicit definitions in the OAIS documentation do not indicate this. OAIS requires information packages to change with the Knowledge Base of the Designated Community [4]. The often described example of what this looks like in practice is a shift in the dominant language (Knowledge Base) of the Designated Community that requires additional translational assistance in archival packages where none was needed before (altered IPs). Another example that stems from the PVWII data is the change in geography over time: Carmen Sandiego games involve chasing 'bad guys' across various geographic locations. The Knowledge Base of the 1985 game player contains the USSR rather than the de-federated former Soviet nations contemporary to the writing of this paper. Maps make for easy pieces of Representation Information to store along with the digital object, all under the OAIS entity *Content Information*.

But when a digital object like a game is seen as imbricated in a complex and ever-changing sociotechnical network, then there are subtler changes to its understandability that are more difficult to document than a dictionary or a map. In the content analysis data of games from PVWII, several Carmen games depict South East Asian countries by employing images of people in conical hats working in rice fields. Today, 31 years after the release of the first *Carmen*, this image still allows game players to identify a certain part of the world, but this knowledge will change rapidly. Water politics and rising sea levels associated with global warming trends mean that large swathes of the Mekong River delta, known as one of the top rice producing and exporting areas in the world, are at risk of being flooded with salt water from the sea. These climate changes threaten to end the farming of rice in these areas: if these trends continue unabated, within a few decades this region will no longer be the center of the rice growing industry. With it will go the cultural association of people in conical hats bent over rice fields as production shifts to Africa, where popular imaginaries suggest different visual markers to note time, place, and occupation. At this point, parts of these video games that rely on tacit knowledge that recognizes images of conical hats and non-descript green fields (this non-description being due largely to technological limitations at the time these games were produced) means that the games can no longer be played: the very behaviors of the digital object break down without enough understanding about the contemporary Knowledge Base of original intended users. And so this situatedness, I would argue, is a significant property in the sense that, without this kind of information, the game is not playable over time even if the bits and computing and rendering environment are preserved. A current presumption of game preservation is that a game, by its nature, is meant to be played, so if it cannot be played, we cannot be said to have preserved a working copy [20].

4.1 PVWII Interviews

Significant properties, as identified in interviews from PVWII, could be located at any point in the layered model developed as part of the grant project.



Application/Experience Layer
Representation of the game
Physical Layer
Console/Computer physical interface to player
Software Support Layer
Operating system, support environment & libraries
Firmware Layer
Built-in instruction functionality for a system
Hardware/Processor Layer
CPU, co-processors, device support hardware

Figure 1: PVWII Layers of a Game

For example, some video games were designed around specific *software support layers,* layer three on the stack, such as the first *Civilization* game designed to work with early Windows operating systems. The functions of the then-novel eponymous windows were incorporated heavily into the game, and constituted a significant property to the developer we spoke with, who mentioned the role this operating system played in the game's development. The Nintendo game *Duck Hunt* notoriously uses a special peripheral *hardware* piece, layer five on the stack. A light gun (as opposed to the normally used d-pad and four-button controller) allowed players to shoot at ducks, as the name of the game implies, and the game is not functional without this piece of equipment. The light gun only functions in conjunction with a cathode ray tube (CRT) television. A CRT monitor might be considered *hardware* or might be considered part of the *physical layer*, layer two in the stack, as part of the physical interface for the player. These twin external hardware dependencies, both of which are essential to a functioning version of *Duck Hunt*, might be considered significant by some Designated Communities.

Yet most significant properties identified by the interview participants in PVWII fell unambiguously under the top, *application* layer of the stack, which is the representation of the game. As a result, I divide the significant properties in this part of the data into three categories, according to where they can be located in relation to the layered model: two of these lie within the top layer of the stack and the third lies outside the stack altogether.

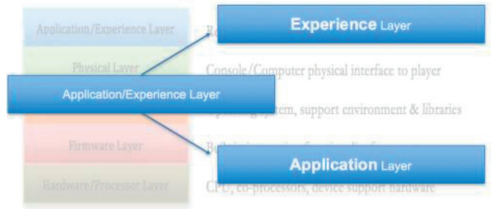

Application/Experience Layer
Physical Layer
Firmware Layer
Hardware/Processor Layer
Console/Computer physical interface to player
tem, support environment & libraries
CPU, co-processors, device support hardware
Experience Layer
Application/Experience Layer
Application Layer

Figure 2. Application and Experience Layers

I firstly break the Application/Experience Layer into two parts: Application and Experience. These encompass many of the significant properties identified by PVWII participants. The application layer includes things like the game code itself, as well as items like jump tables for early Mario Brothers games or historical statistical mortality data that determined how likely a player was to die when playing Oregon Trail. The experiential layer encompasses the surface and affective experiences of playing the game: the fact that Carmen Sandiego is only kind of a one-player game despite its single avatar because of its situation in public schools, for example. I separate Application and Experience because I argue that they are not necessarily related. To be sure, the original code and a computing environment were necessary to manifest the original playing experience. But to recreate the mnemonic experience [31], to give an authentic representation of the experience of play, the original code is no longer necessary. PVWII investigators posed the question to game designers: how important is the original code if you can generate the same surface appearances and behaviors with a different backend? Most responded that they were not wedded to the original code, but more so to the experience of play. Some noted that the original code itself was 'poor', often due to time constraints. These two things can exist separately: because it is possible to save 1s and 0s and even consoles and media without saving the experience and it is possible to recreate the experience without the 1s and 0s, I separate this layer into two discrete layers.

Finally, I also argue that some kinds of significance, as described by PVWII participants, lie outside the stack altogether—that is to say, *external significant properties* cannot be found in the code or environment. These include significant properties like those I term *relationally significant*. PVWII investigators asked participants to name their favorite game franchise and to explain what made it so important. One point of significance that was mentioned was a game that was 'leaps and bounds' ahead of its predecessors and contemporaries. I term this *relationally significant* because understanding this statement about what makes a game important requires placing it in context, almost like archival bond, with other games of its time. To understand this property of a game does not require a playable copy, although it might entail placing a playable copy up against playable copies of its contemporaries to demonstrates its advancements. But there are other ways to represent and benchmark this: for example, placing visuals from saved game files, videos of play, or machinima in relation to visuals of its contemporaries. The constant feature here is that is impossible to understand the "advancedness" of a game by looking at the game itself: it has to be seen in relation to other things.

PVWII interviews also raised other external significant properties of game play. For Carmen Sandiego and Oregon Trail, two franchises that are often termed *edutainment* games to the chagrin of their producers, interviewees expressed that understanding them in the educational context of the mid-1980s is important to understanding the experience of play. Like the tangible difference between playing the original arcade *Donkey Kong* and playing a game on a home console (one designed to be short to eat money, the other designed to be long to engender customer loyalty to a product), there is a tangible difference between playing Carmen Sandiego solo at home today with Google, versus playing it in its original environment: several kids around a tiny and expensive computer in a school, with one person at the keyboard and another working the accompanying encyclopedia. One interview participant who worked on programming for Carmen Sandiego said that seeing these games in context was how he envisioned ideal preservation for his games, while acknowledging the difficulty in manifesting something like the mnemonic impressions of a particular time and place.

4.2 OAIS Interviews

While coding the interviews with OAIS practitioners, authors, and scholars, there were only three instances in which the specific term *significant properties* was brought up by my interviewees, and this makes sense given that the dominant theme of these interviews was OAIS and the fact that significant properties is not an OAIS term. The explicit significant property instances echo the dominance of the OAIS authors in shaping how people within the realm of digital preservation continue to respond to and understand their work in relation to OAIS. In one instance, I asked an interview participant about significant properties specifically because I knew this participant had made public statements about them in relation to OAIS in the past. In this particular discussion, the interviewee mentioned significant properties in relation to enrolling analog professionals within libraries in digital work. The interview subject said:

> *"...we have more analog material... and lots of people were trained to deal with analog material and fewer people are trained to deal with digital material. And as you can't just give them the sack [laughs], you need to deal with them, you train them or whatever, so that takes a long time and I think that's one of the problems all main libraries are dealing now with that they have staff that's not quite prepared for digital material. So that the thinking about OAIS starts within a... small group of people... and we tell them 'I think you should interpret it like this or like that' and what you don't see is that they try to translate it to their analog environment and sometimes that does not work because it's digital. So it's difficult to translate I think, although the model itself is very clear, I think it's rather straight forward, but when you go the significant properties, well, endless discussions."*

This interview participant, someone who is both a library practitioner and actively involved in OAIS revisions and related standards, describes OAIS as "relatively simple". In this case, the designation of simplicity is meant, as much as anything, to indicate how not simple the concept of *significant properties* is. The situation in which she is working is already a fraught one to some extent: the library has a large analog collection and many analog employees, and moving into the digital space requires people to learn new skills. And it is under this umbrella discussion about employees who work with analog materials, who cannot make analogies between their previous work and their digital futures, and who struggle with a "simple" model often because they "only read the first 80 pages [of OAIS]" according to the same interviewee, that the subject of significant properties arises. As suggested elsewhere, this notion comes from library and archive traditions, and therefore clashes with data and systems design origins that dominate the construction of OAIS. This is the unresolved tension a reviewer noted in response to an article I submitted on the subject to a major preservation-oriented conference. And perhaps it is the perception by OAIS authors that significant properties come from libraries and archives that predicates its continued exclusion from OAIS.

A second mention of significant properties in relation to OAIS came from a US-based data scientist who said:

> *"I mean, if OAIS didn't exist, you know, people would still need to preserve things and they would come up with some other framework, and obviously it would be not exactly the same as OAIS. It would probably have a lot of the same ideas in it. There were, you know, obviously... concepts that I used before I ever saw OAIS, but when I saw it, I thought, "Oh, yeah, this maps to this in OAIS." And OAIS has concepts in it from earlier versions of OAIS that aren't the same anymore like format migration isn't called format migration anymore, it's called transformation. And significant properties are now like transformational information property, you know, and things like that."*

This suggests a familiarity with the process of OAIS creation and revisions, such that this person is aware of the fact that Transformational Information Properties are the official term meant to deal with significant properties. This interview subject speaks from a place of privilege: as a science data scholar, this person was already familiar with the type of terminology that is contained within OAIS, and is happily fluent in its lingua franca. In fact, of all my participants, this one had the fewest complaints about OAIS, expressing most answers in form similar to the quote above.

The comment by US-based data scientist about the relationship between significant properties and Transformational Information Properties is a common misconception, if it can be called that. It may simply be a casual simplification. While Transformational Information Properties are meant to encompass some aspects of significant properties, they are not a replacement. Defined in the 2012 revisions [4] as an:

> *"[i]nformation [p]roperty the preservation of the value of which is regarded as being necessary but not sufficient to verify that any Non-Reversible Transformation has adequately preserved information content. This could be important as contributing to evidence about Authenticity. Such an Information Property is dependent upon specific Representation Information, including Semantic Information, to denote how it is encoded and what it means. (The term 'significant property', which has various definitions in the literature, is sometimes used in a way that is consistent with its being a Transformational Information Property)."*

It is stated quite clearly that this definition is meant to cover only some definitions of significant properties. Depending on the definition of significant properties one employs from among the myriad ones in existence, some of these properties are contained within entities that predate the 2012 revisions, including within the Digital Object itself as well as in places like the Preservation Description Information entity, without necessary reference to a Non-Reversible Transformation.

These are two distinctly interesting explicit mentions of significant properties from the OAIS interview data. The more populous node, however, was *implicit significant properties*. I applied this label to any discussions wherein an interviewee mentioned some aspect of a digital object without which that object would not be understandable, functional, authentic, or worth preserving; in other words, specific values labeled by the participants with any of the descriptors from the myriad definitions of significant properties at the outset of the paper. The findings from this node within the OAIS data include a number of references that echo the PVWII data. One practitioner mentioned a concern about the dependency on outside objects for understandability, in particular external technologies. This US-based museum practitioner also said:

> *"Yeah, like Windows '95, we need a place to track that information and because there is a many-to-many relationship there, it makes sense to record that in a structured way where we have some kind of master record of all these technologies."*

This comment was in reference to the difficulty of creating mutable AIPs within the software programs the institution uses for documenting art records. The substance of the comment mirrors discussions with video game creators who referenced the significance of the role of the operating system, coincidentally also Windows '95, in the creation of a title within an iconic video game franchise.

Likewise, the experiential aspects of digital objects also arose in the OAIS interviews. One participant, a digital preservation manager at a private US university, said:

> *"...Maybe we need to be more clear about it's not just about providing [access] to the files, it's about providing an experience... I mean, I like to think about it as being able to present the same content to the user...we could have documented that content, regardless of the experience through which they receive that content, even if the content is an experience... I don't know. It's complicated...And I also feel like... just in general... there's so much interaction, and the experience of being able to work and build, something like that."*

This is not to say that engaging with analog materials is not experiential: indeed, reading a paper book is an experience, and reading a Dickens novel as a set of serialized chapters over the course of months is not quite the same experience as reading the entire work at once when it has been collected into a single volume. But in this case, the interviewee is expressing something fundamental about the interactivity of many types of digital content. I take this ethos to be the same one that motivates the response on the part of video game programmers that the look and feel and even social experience of playing a game may be more important to preserve than the code. This is precisely the difficulty that preservationists face with dynamic and interactive content. Cases like video games offer heuristics that demonstrate one of the chief difficulties in the realm of preservation: it it very hard to predict the future. The difficulty is how to demonstrate, through the Dissemination Information Package (DIP), the temporal, spatial, and social aspects of content.

Conveying this information back to users is a function of multiple entities within the OAIS information model. First, an Archive must store sufficient information within its Archival Information Packages (AIP) to to be able to convey external significant properties or to change Information Packages to match changing Knowledge Bases: this includes something like the OAIS/FRBR mapping constructed as a result of PVWI wherein the model suggests linking to an outside source for Context and Provenance information [21]. Perhaps in a case like the ‘leaps and bounds’ advancement of a particular video game, the AIP would contain not only the game, but also references to popular articles, industry reviews, and fan content. The digital preservation manager at a private US university quoted above describing the interactive nature of technology also mentions the practice of documenting the experiences of users. For very complicated media that is one of the few (perhaps the only) options at this point in time. Another interview participant, a researcher at a European national archive, said:

> *“I looked at technological hardware preservation. I looked at simulation—yeah, migration and emulation then documentation. Documentation is kind of like a separate thing but I felt because so many of these other things there are so many reasons why we can’t really do that yet. I feel like documentation is basically what we’re left with.”*

The second entity that is implicated in conveying mnemonic experience is the DIP. These types of experiential significant properties require creative work through DIPs to deliver authentic experiences to Consumers. The DIP is one of the more poorly defined entities within OAIS in large part, as one interview subject said, because it requires a prediction of the future. When Knowledge Bases change and people no longer understand how a d-pad works, the DIP for a Nintendo game has to go beyond simply providing a working console and cartridge to a Consumer. There is recent work that provides formal modeling of DIPs displayed as a set of services and exchanges with Consumers [12] and this work acknowledges the need for DIPs to change according to queries by Consumers; it suggests tracking different DIPs and the queries the spurred their generation and potentially adding them back into the AIP using the PREMIS standard for documentation; the most recent version of PREMIS even allows for the documentation of environments as their own objects, a move that recognizes that environments may be significant in the preservation of content beyond the bits themselves [6]. But even though PREMIS is a more specific and prescriptive standard that follows OAIS, it does not and perhaps cannot help to address what will need to be somewhat imaginative solutions for conveying the experiences of interactive and dynamic digital content. This entire concern is imbricated in the complexity of Designated Communities and Knowledge Bases. Archives are supposed to track Knowledge Bases and update content when Knowledge Bases change. This is a difficult task, not only because there are no current guidelines that deal specifically with this[1], but also because change is both a hard thing to notice in the moment and a more difficult thing to document after the moment has passed. That there is no one solution is part of what makes this kind of thing hard to standardize; that there should be guidelines anyway is probably obvious given the complexity of the task.

There are also ways in which it may be possible to overstate the difficulty of the digital preservation task: it may be that at this particular juncture, the preservation of surface and affective experience is not possible, particularly not at scale. One interview participant, a senior digital preservation consultant at a boutique US firm, noted that these preoccupations can serve to paralyze the field in such a way that getting to grips with what should be relatively simple tasks like bit-level preservation still have not been definitively addressed:

> *“Yeah, I think it actually… and this isn't OAIS's fault, it's just I think this field has suffered from -- in my opinion, it has suffered from too much fixation on those kinds of issues and not just doing the absolute minimum to get you to a point to have a future opportunity to visit those questions when the need really arises. We don't even have good bit preservation nailed down, and that should be very easy. It's really simple, it's dumb, just do it, and stop talking about it, please. I'm so tired of it.”*

This participant also noted that concerns about significant properties are more challenging for some kinds of content that others. For audiovisual materials, she argued: “Watch it and listen to it, and look at it.” Another participant, a senior special collections archivist at a private US university, said, “So, for us to be able to push [a digital object] into something where we have, you know, huge, huge disk space, and to be able to say well, at least you know, it's safe, the original is safe. I would think that would be like a big plus to people, just to be able to provide that as a service for their materials.” Keeping the 1s and 0s safe is a most basic requirement, and this might be seen as sufficiently significant in many cases, particularly if this is explicitly stated in users and donor agreements. Yet at the same time, multiple people have pointed out, including the authors of OAIS that I spoke to, that 1s and 0s alone are rarely sufficient, particularly when longer time scales are involved.

5. DISCUSSION

The previous section detailed some of the significant properties that arose in conversations with game programmers and OAIS practitioners and scholars. Here, I will demonstrate what maps well to the existing OAIS entities and what works less well. The figure below is an image from OAIS that details the contents of the AIP. I have highlighted in purple the entities wherein some significant properties could be located and I speak about some of these in the examples that follow.

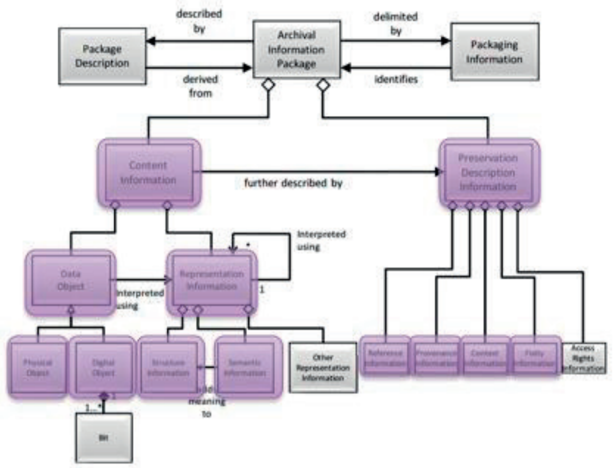


Figure 3. Significance in OAIS

Some significant properties fit well within the Content Information entity in the AIP model. Content Information includes the Data Object itself, which can be comprised of both Digital Object(s) (bits) and Physical Objects. Source code of games fits here as do some physical ephemera essential for use, like *Carmen Sandiego*’s analog copy protection *World Encyclopedia*. Ephemera can also be documented as a separate object and related to the digital data via the Context Information entity.

Access software and, by extension, access hardware may be documented as part of the Data Object itself or as Structural Representation Information. Changes in the Designated Community’s Knowledge Base may be documented as Semantic Information, although there are limits. Including software as part of the digital object itself is something that OAIS does not do very well yet, according to some practitioners. One of the interview subjects has argued vociferously and publicly for its inclusion as part of the object itself in the 2017 revisions. Semantic Information can document significant properties like a language shift from English to Chinese, for example. Preservation professionals interviewed disputed whether or not it is the role of the repository to document changes in common knowledge, such as geographical names and borders or popular imaginaries in the case of Carmen Sandiego.

Sometimes what is significant about a game is its relationship to other games. One game programmer said, “*Doom*, for example, it made some of these huge graphics and texturing leaps and bounds, [these were] obviously… a product of its time.” “Leaps and bounds” progress in one game necessarily relates it to a history wherein a game was markedly different that its contemporaries, as noted previously. Another significant property noted by interviewees is the relationship of a particular title to a larger franchise, for example a particular release of *Civilization* in relation to all versions. This was stated explicitly but is also tacitly implied when participants spoke about franchise games by collapsing an entire series into a single sociotechnical entity, saying things like, “Civilization is one of my really favorite games of all time,” as opposed to naming a particular version or release of Civilization. In OAIS, this relationality can be mapped as Context Information within the Preservation Description Information entity. What is meant by Context Information is unclear to some interview participants; its description in the OAIS literature is similar to archival bond. Therefore, a repository can only express this Significant Property as Context Information if it holds enough games to demonstrate how a particular game relates to others.

Many interviewees acknowledged that preserving the affective and social aspects of games is a most challenging task. Playing games in arcades is a fundamentally different experience than playing at home; these locations impact game design, for example the simplicity of original *Donkey Kong* versus the deeper interaction of *Super Mario Brothers*. Creators and players describe the school-setting of the earliest *Oregon Trail* and *Carmen Sandiego* titles as a significant property. The need to understand the time and place in which a game was made and/or played might be easiest to understand with a game like *September 12th*, a news game predicated on the events of September 11, 2001. The twin difficulties are encapsulated by two quotes from different game developers. The first, a contemporary developer working in a US game studio, said, “…it’s hard to differentiate between what is like your nostalgia and what is sort of useful, right?” A second quote, from a developer of a game series that is no longer in production, said, “So you really have to sort of capture the essence of the time. Now I don’t obviously have a good answer for that, but somebody should think about it.”

These Significant Properties do not fit well within OAIS. This may be because documenting this type of information in relation to a particular object has not always been seen as the province of the archive itself. In some cases, the preservation of some non-code significant properties of a game is more desirable than preserving working code itself: a video of game play, a textual narrative of a walk-through may better capture the experience than working copies of obsolete technology. In fact, these expanded descriptions of what might be significant about a game challenge the very assumption that a baseline for a game’s authentic preservation is its functionality.

6. CONCLUSIONS

Some significant properties, as suggested by interviewees in OAIS research and PVWII respectively, fit well within the existing OAIS entities. For others, one could argue for their inclusion within existing entities although it may mean stretching the capacity and meaning of these entities beyond what was envisioned by the designers of OAIS. This latter is not to indicate that such actions would be wrong: indeed, it is the role of a reference model to inform things in the future which likely entails moving into spaces the original authors could not envision.

Data from PVWII suggest that social and affective attributes of games are considered significant by designers and players. These significant properties are largely expressed as relational properties: they obtain in relation to objects, events, spaces, and times outside the object and often outside the archive or repository. These relationships are also nuanced in nature: certain properties are more important than others, or are only important in certain cases (for example, to particular Designated Communities). In fact, the situatedness of significant properties suggests that, for popular content like video games, the notion of Designated Communities is too vague and it is more important to think about archived objects in the context of Ranganathan’s [24] third law: every [digital object] its [user]. Video games serve as an excellent case study for this type of research precisely because they are complex technological objects but also because their heterogeneous users offer up a complicated sociotechnical network within which to understand something like significance. But these findings are not specific to video games: rather, the case study serves to bring to the fore issues that are already present in long-standing preservation practices for analog materials and that are currently under debate for digital materials such as scholarly data, media art, web archiving, and the nebulous notion of digital archives more broadly.

Some significant properties identified from within the PVWII data fit within OAIS, such as semantic and environment information; others will require either new metrics or changes to the existing standard, like affective and relational values. These findings are echoed by similar comments from OAIS interview subjects, and this is all the more pertinent given both the variety of participants in this latter study and the fact that the conversations I had with them were very different in nature and subjects from the PVWII interviews. The similarities between the two datasets, PVWII and OAIS experiences, speak to the salience of these themes beyond the theme of video games and within wider digital preservation discourse.

What was surprising about this project was just how much data I struggled to map to OAIS: my original hypothesis when I began this mapping project was that all Significant Properties should fit within OAIS, given its commitment to changing Knowledge Bases over time. For example, the process of documenting context is nothing more than moving additionally pre-inscribed affordances of a digital object into the circumscribed setting of the Archive. In the language of OAIS, adding information from the Knowledge Base of the Designated Community to the AIP as additional documentation is taking what is normally afforded to the stakeholders forming the Designated Community and pulling it into the AIP. This further is an extension, and not necessarily incommensurate, with earlier work done on significance in OAIS [13] and work on significant characteristics [7]. What this paper suggests is merely an

[1] Although one interview subject suggested the outcomes of the SCAPE project [28], suggesting “a lot of the idea in the SCAPE approach of preservation monitoring and planning is predicated on evolution of and instruction of the designated community in technology, in semantics, in usage, in requirements.”

extension of an on-going balancing act, of finding the line for sufficiency in deciding how much to document: this is precisely why I call for the creation of metrics to help drawing these artificial boundaries so that this work can be made machine-actionable for digital preservation at scale. My conclusion is, therefore, that all significant properties do not fit within existing OAIS entities and I echo the calls of other preservation scholars that changes are needed in the ways in which we think about the responsibilities of repositories, especially given the potential for distributed digital preservation in linked data environments. Additionally, I posit that these difficulties will be exacerbated in areas where OAIS already does not work as well. A couple interview participants noted that the scripts within OAIS presume a level of infrastructure. While Seles [29] demonstrates how this plays out in situations where Archives are located in geographical regions where the legal, electrical, and network infrastructure are missing, some of my interview participants pointed out that, even in wealthy first world contexts, institutions wherein preservation is not a primary function will lack many of the structures presupposed by OAIS.

In this paper, I do not tackle the breadth of descriptions or definitions about what significance actually means, whether characteristics, properties, or anything else. In fact, this work encompasses many of the definitions from digital preservation literature. Instead, I locate claims that significant properties are situated and sometimes outside the digital object and it computing environment within a growing body of archival science literature that speaks to the situatedness of archival content and what is needed to contextualize it [18]. The juridical and legal undercurrents of archival conceptions of authenticity are balanced by work in practice, where archivists understand that evidence, for example, aids in interpretations of the world [5] and that archives may have the role of preserving mnemonic devices in addition to evidence [31]. What is necessary is for digital preservationists to decide whether what is wanted is particular bits of information or impressions of the past.

7. ACKNOWLEDGEMENTS

I wish to thank all the investigative partners from Preserving Virtual Worlds II, including colleagues at the University of Illinois, University of Maryland, Rochester Institute of Technology, and Stanford University. PVWII was supported by a grant from the Institute of Museum and Library Services; additional research was supported by the Graduate School of Library and Information Science at the University of Illinois, Urbana-Champaign. I also acknowledge Jerome McDonough, Gail Ferguson, and Nikki Weickum for their feedback on earlier drafts.

8. REFERENCES

[1] Bates, M. (2015). The Information Professions: Knowledge, Memory, Heritage. *Information Research: An International Electronic Journal*, *20*(1)

[2] Bradley, K. (Summer 2007). Defining Digital Sustainability. Library Trends 56(1): 148-163.

[3] Cargill, C. (1997). *Open Systems Standardization: A Business Approach*. Upper Saddle River, NJ: Prentice Hall.

[4] Consultative Committee for Space Data Systems (June 2012). *Reference Model for an Open Archival Information System (OAIS).* CCSDS 650.0.M-2, Magenta Book. Issue 2.

[5] Caswell, M., & Gilliland, A. (2015). False promise and new hope: dead perpetrators, imagined documents and emergent archival evidence. *International Journal Of Human Rights*, *19*(5), 615-627. doi:10.1080/13642987.2015.1032263

[6] Dappert, A. (2015). Digital Preservation Metadata and Improvements to PREMIS in Version 3.0. Webinar.

[7] Dappert, A., & Farquhar, A. (2009). Significance is in the eye of the stakeholder. In *Research and Advanced Technology for Digital Libraries* (pp. 297-308). Springer Berlin Heidelberg.

[8] Dappert, A., Peyraud, S., Delve, J., Chou, C. (2012). Describing Digital Object Environments in PREMIS. In: iPRES 2012 - 9th International Conference on Preservation of Digital Objects, October 3-6, Toronto, Canada.

[9] Delve, J. (2013). Introduction to POCOS e-Book 3: Preserving Gaming Environments and Virtual Worlds In The Preservation of Complex Objects: Volume 3. Gaming Environments and Virtual Worlds. JISC.

[10] Duranti, L. (1994). "The concept of appraisal and archival theory" in The American Archivist 57, pp. 328-344.

[11] Duranti, L. (2010). "The appraisal of digital records: the return of diplomatics as a forensic discipline." International Conference on Surviving the Digital Wild Frontier. Singapore, Singapore.

[12] E-ARK. (2015). E-ARK DIP Draft Specification. D5.2, Revision 4

[13] Giaretta, D., Matthews, B., Bicarregui, J., Lambert, S., Guercia, M., Michetti, G., Sawyer, D. (2009). Significant Properties, Authenticity, Provenance, Representation Information and OAIS Information. iPRES 2009: the Sixth International Conference on Preservation of Digital Objects. California Digital Library, UC Office of the President

[14] Hedstrom, M. and Lee, C.A. (2002). “Significant properties of digital objects: definitions, applications, implications”, Proceedings of the DLM-Forum 2002. Retrieved from http://ec.europa.eu/transparency/archival_policy/dlm_forum/doc/dlm-proceed2002.pdf

[15] KEEP (Keeping Emulation Environments Portable). http://www.keep-project.eu/ezpub2/index.php

[16] Knight, G. (2008). Framework for the definition of significant properties. Retrieved from http://www.significantproperties.org.uk/documents/wp33-propertiesreport-v1.pdf

[17] Lee, C. (2005). “Defining Digital Preservation Work: A Case Study of the Development of the Reference Model for an Open Archival Information System.” Doctoral Dissertation. The School of Information, University of Michigan, Ann Arbor.

[18] MacNeil, H. and Mak, B. (2007). "Constructions Of Authenticity." *Library Trends* 56.1, 26-52.

[19] Matthews, B., McIlwrath, B., Giaretta, D., & Conway, E. (2008). The significant properties of software: A study. *JISC report, March*.

[20] McDonough, J. (2013). A Tangled Web: Metadata and Problems in Game Preservation In The Preservation of Complex Objects: Volume 3. Gaming Environments and Virtual Worlds. JISC.

[21] McDonough, J., Olendorf, R., Kirschenbaum, M., Kraus, K., Reside, D., Donahue, R., Phelps, A., Egert, C., Lowood, H., & Rojo, S. (2010). Preserving Virtual Worlds Final Report. Retrieved from IDEALS online 2011-11-30.

[22] Meghini, C. (2013) Data preservation. Data Science Journal Volume 12, 10 August 2013, pp. GRDI51-GRDI57

[23] Pinchbeck, D., Anderson, D., Delve, J., Ciuffreda, A., Otemu, G., Lange, A. (2009). Emulation as a Strategy for the Preservation of Games: The KEEP Project

[24] Ranganathan, S. R. (1931). The five laws of library science.

[25] Rosenthal, D. (2015). The case for a revision of OAIS. Digital Preservation Coalition OAIS Wiki. Available at: http://wiki.dpconline.org/index.php?title=The_case_for_a_revision_of_OAIS

[26] Sacchi, S., Wickett, K., Renear, A., and Dubin, D. (2011). A Framework for Applying the Concept of Significant Properties to Datasets. In Proceedings of ASIS&T 2011.

[27] Schwartz, J. & Cook, T. (2002). "Archives, Records, and Power: The Making of Modern Memory," *Archival Science*, 2:1/2: 1-19.

[28] SCAPE. Scalable Preservation Environments: http://scape-project.eu/

[29] Seles, A. (2016). The Transferability of Trusted Digital Repository Standards to an East African Context. Doctoral Thesis. Department of Information Studies, University College London. London, UK.

[30] Sierman, B. (2012). “OAIS 2012 Update” in Digital Preservation Seeds Blog accessible at http://digitalpreservation.nl/seeds/oais-2012-update/

[31] Smith, A. (Summer 2003). Authenticity and Affect: When is a Watch Not a Watch? Library Trends 52(1), pp. 172-82.

[32] Webb, C., Pearson, D., and Koerbin, P. (2013). ‘Oh, you wanted us to preserve that?!’ Statements of Preservation Intent for the National Library of Australia’s Digital Collections*. D-Lib Magazine 19:1/2* DOI= doi:10.1045/january2013-webb

[33] Yeo, G. (2010). "‘Nothing is the same as something else’: significant properties and notions of identity and originality." Archival Science 10, no. 2: 85-116.

Database Preservation Toolkit

A relational database conversion and normalization tool

Bruno Ferreira
KEEP SOLUTIONS
Rua Rosalvo de Almeida 5
4710 Braga, Portugal
bferreira@keep.pt

Luís Faria
KEEP SOLUTIONS
Rua Rosalvo de Almeida 5
4710 Braga, Portugal
lfaria@keep.pt

José Carlos Ramalho
University of Minho
4710 Braga, Portugal
jcr@di.uminho.pt

Miguel Ferreira
KEEP SOLUTIONS
Rua Rosalvo de Almeida 5
4710 Braga, Portugal
mferreira@keep.pt

ABSTRACT

The Database Preservation Toolkit is a software that automates the migration of a relational database to the second version of the Software Independent Archiving of Relational Databases format. This flexible tool supports the currently most popular Relational Database Management Systems and can also convert a preserved database back to a Database Management System, allowing for some specific usage scenarios in an archival context. The conversion of databases between different formats, whilst retaining the databases' significant properties, poses a number of interesting implementation issues, which are described along with their current solutions.

To complement the conversion software, the Database Visualization Toolkit is introduced as a software that allows access to preserved databases, enabling a consumer to quickly search and explore a database without knowing any query language. The viewer is capable of handling big databases and promptly present search and filter results on millions of records.

This paper describes the features of both tools and the methods used to pilot them in the context of the European Archival Records and Knowledge Preservation project on several European national archives.

Keywords

Preservation; Archive; Relational Database; Migration; Access; SIARD

1. INTRODUCTION

Databases are one of the main technologies that support information assets of organizations. They are designed to store, organize and explore digital information, becoming such a fundamental part of information systems that most would not be able to function without them [5]. Very often, the information they contain is irreplaceable or prohibitively expensive to reacquire, making the preservation of databases a serious concern.

The Database Management System (DBMS) is the software that manages and controls access to databases, which can be described as a collection of related data. These two intrinsically related technologies function together to perform tasks such as information storage and retrieval, data transformation and validation, privilege management and even the enforcement of important business constraints. The most popular databases are based on the relational model[1] proposed by Codd. [5, 4]

The migration of the relational database information into a format well suited for long-term preservation is one of the most accepted strategies to preserve relational databases. This strategy consists in exporting the information of the relational database, including descriptive, structural and behavioural information, and content, to a format suitable for long-term preservation. Such format should be able to maintain all significant properties of the original database, whilst being widely supported by the community and hopefully based on international open standards [7]. Few formats fit this criteria, being the SIARD format one of the main contenders.

The Software Independent Archiving of Relational Databases (SIARD) format was developed by the Swiss Federal Archives and was especially designed to be used as a format to preserve relational databases. Its second version, SIARD 2, retains the (most commonly agreed upon) database significant properties and is based on international open standards, including Unicode (ISO 10646), XML (ISO 19503), SQL:2008 (ISO 9075), URI (RFC 1738), and the ZIP file format. [6, 8, 9, 11, 14, 15]

The manual creation of SIARD files is impractical, therefore an automatic conversion system was developed – the Database Preservation Toolkit (DBPTK). This software can be used to create SIARD files from relational databases in various DBMSes, providing an unified method to convert databases to a database agnostic format that is able to retain the significant properties of the source database. The software uses XML Schema Definition capabilities present in the SIARD format to validate the archived data and can also be used to convert the preserved database back to a DBMS.

The digital preservation process is not complete if the archived information cannot be accessed. To access and explore digitally preserved databases, the Database Visualization Toolkit (DBVTK) is being developed. This software can load databases in the SIARD format and display their descriptive, structural and behavioural information and content. The viewer also provides the functionality to search and filter the database content as well as export search results.

[1] According to the ranking at http://db-engines.com/en/ranking (accessed on Apr. 2016), where 7 of the top 10 DBMS use the relational model

```
/ ............................................... zip file root
├─ header/ .................... folder for database metadata
│  ├─ metadata.xml
│  ├─ metadata.xsd
│  └─ version/
│     └─ 2.0/ ............ empty folder signalling version 2
└─ content/ ..................... folder for database content
   └─ schemaM/ ........... M is an integer starting with 0
      └─ tableN/ .......... N is an integer starting with 0
         ├─ tableN.xml ......... same N used in tableN/
         └─ tableN.xsd ......... same N used in tableN/
```

Figure 1: Basic SIARD 2 directory structure.

2. SIARD 2

The second version of the SIARD format emerged from lessons learnt by creators and users of database preservation formats. The SIARD format was originally developed by the Swiss Federal Archives in 2007 and is being used by many archives worldwide. In 2013, the SIARD format became a Swiss E-Government Standard (eCH-0165). The SIARD-DK is a variation of the SIARD format created by the Danish National Archives to fit their specific needs. The Database Markup Language (DBML) format, created at University of Minho, was used by the Repository of Authentic Digital Objects (RODA)[2] software to preserve databases at the Portuguese National Archives[3]. The Archival Data Description Markup Language (ADDML) is the format used by the Norwegian National Archives to describe collections of data files. [3, 12, 13, 1]

The SIARD 2 format, in its most basic form, consists of a ZIP file that contains a hierarchy of folders and files of XML and XSD (XML Schema) format, illustrated in figure 1. The XML files inside the SIARD file hold database metadata information and contents.

The `metadata.xml` file contains database description information, such as the database name, description and archival date, the archivist name and contact, the institution or person responsible for the data; database structural information, including schemas, tables and data types; and behavioural information like keys, views and triggers. Such information is useful not only to document the database but also to allow the reliable export of its structure on a different DBMS.

The `table`*N*`.xml` files correspond to each of the database tables and hold the content of the rows and cells from that table. All XML files are accompanied by a corresponding XML Schema file, that can be used to validate the structure and contents of the XML files.

The SIARD format includes advanced features such as the support for Large Objects (LOBs)[4], and ZIP compression using the *deflate* method [15]. SIARD 2 brings many improvements over the original SIARD and other database preservation formats, mainly the support for SQL:2008 standard and data types, including arrays and user defined types; the strict validation rules present in XML Schema files to enforce valid XML structure and contents; and allowing LOBs to be saved in the `table`*N*`.xml` file, saved as files in a folder inside the SIARD, or saved as files in a location outside the SIARD file. Furthermore, the SIARD 2 specification allows LOB files to be saved in multiple locations or storage devices outside the SIARD file, increasing support for databases which contain large amounts of LOBs.

3. DATABASE PRESERVATION TOOLKIT

The DBPTK is an open-source project[5] that can be executed in multiple operating systems and run in the command-line. It allows the conversion between database formats, including connection to live Relational Database Management Systems, for preservation purposes. The toolkit allows extraction of information from live or backed-up databases into preservation formats such as SIARD 2. Also, it can import back into a live DBMS, to provide the full DBMS functionality, such as SQL[6] querying, on the preserved database.

This tool was part of the RODA project and has since been released as a project on its own due to the increasing interest on this particular feature. It is currently being developed in the context of the European Archival Records and Knowledge Preservation (E-ARK) project together with the second version of the SIARD preservation format – SIARD 2.

The DBPTK uses a modular approach, allowing the combination of an import module and an export module to enable the conversion between database formats. The import module is responsible for retrieving the database information (metadata and data), whilst the export module transcribes the database information to a target database format. Each module supports the reading or writing of a particular database format or DBMS and functions independently, making it easy to plug in new modules to add support for more DBMS and database formats. The conversion functionality is provided by the composition of data import with data export.

Currently supported DBMSes include Oracle, MySQL, PostgreSQL, Microsoft SQL Server and Microsoft Access. All of these support import and export, except Microsoft Access where only import is available, i.e. conversion *from* Microsoft Access is possible, but conversion *to* Microsoft Access is not. All these modules use the Java Database Connectivity (JDBC) modules as a generic starting point, and then deviate as much as needed to account for functionality specific to each DBMS.

The base JDBC import and export modules, given the correct configurations and dependencies, may enable con-

[2] Current version of RODA is available at http://www.roda-community.org/

[3] Direcção-Geral do Livro, dos Arquivos e das Bibliotecas

[4] Large Objects, usually stored in binary format in databases. Can also be referred to as BLOBs or CLOBs if they contain binary or character data, respectively.

[5] Software and documentation available at http://www.database-preservation.com

[6] SQL: Structured Query Language

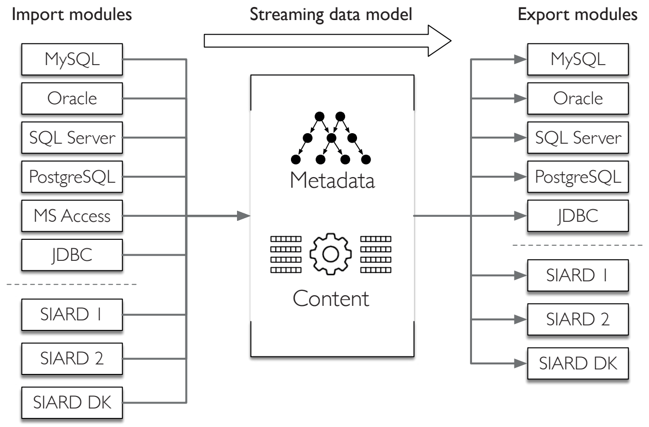


Figure 2: Application architecture overview

version to or from any DBMS. Being DBMS-agnostic, this technology can be used to connect to a wide range of database products, however some specific features (e.g. data types) may not be supported and may hinder a fully successful conversion.

The SIARD modules are an essential part of DBPTK, being used to allow the conversion of databases to and from this database preservation format. The SIARD export modules allow filtering out some database tables, as well as exporting contents from views as if they were database tables.

Attending to security and data privacy concerns, modules default to using secure (encrypted) connections to DBMSes if such a connection is supported by the DBMS.

Figure 2 depicts an overview of the information flow in the application, with import modules as information providers, extracting the information from a source and mapping it into an internal application model; and export modules implementing the inverse process, by mapping information from the internal model to the target DBMS. This mapping may be specific for each DBMS module, as most DBMSes have specific or incompatible features.

In the first phase of the conversion, the database metadata (descriptive, structural and behavioural information) is fetched by the import module and transcribed to the target Database Management System or format by the export module. This phase is followed by the conversion of database contents. Using streaming and optimizing interactions with the database, the contents are converted, record by record, with low computing resource requirements. Finally, system resources are released, concluding the execution. While this is a common overview of a typical execution, specific modules can slightly diverge from this approach to improve performance and error handling.

The conversion is prioritized by firstly converting the database content information without loss, secondly trying to keep the database structural metadata identical to the original database, and thirdly attempting to translate the database behavioural metadata to the target database. In practical terms, this means that in cases where the target DBMS does not support the data type used in the original database, an equivalent or less restrictive data type is used; this changes the database structure metadata, but avoids database content information losses. The database behavioural information is the last priority in the conversion because it is prone to failure (with a warning) due to source DBMSes that do not check the invariants and constraints imposed by behaviour like primary and foreign keys, or views which have DBMS-specific and untranslatable queries, not supported in the target DBMS.

Figure 4 introduces an overview of a database structure as a hierarchy. As most database structures fit this structure entirely or partially, it is used by all modules. However, there are some database systems, e.g. MySQL, that do not fit this structure entirely, as they have no schemas. In these cases, all the information that would be accommodated in a schema is moved up to the database component, resulting in a slightly different component that performs as both a database and a single schema, depicted in figure 5. DBPTK import modules work around this issue by treating the schema-less database as if it were a database containing a single schema, moving any tables, views, routines and user defined types to this schema.

Most DBMSes implement SQL with slight deviations from the SQL standard. These derived query languages are commonly referred to as SQL flavours and make it difficult to create a set of queries compatible with the majority of DBMSes. To create queries, there is a query generator, based on the SQL standard, serving as a base for a few flavour-specific query generators. The import and export modules use the most specialized SQL generator considering the DBMS SQL flavour, guaranteeing equivalent functionality across different DBMSes.

SQL flavours often include new SQL data types or alias to standard SQL data types, but internal data types used in DBPTK are based on SQL standard data types. During the database conversion process, the import module maps the data types to appropriate internal data types, and the export module does the inverse process, by mapping the internal data types to data types supported by the target Database Management System or format. The aforementioned process is observable in figure 3.

Most DBMS implementation specific SQL types are automatically converted to standard SQL types as they are obtained by the import modules, but there are a few cases that need to be handled specially for each DBMS. An example of such case is the `YEAR` MySQL data type[7], depicted in figure 3, which the import module first perceives as representing a date, but is in fact a 4 digit numeric type (corresponding to the SQL:2008 standard type "`NUMERIC(4)`"). Since PostgreSQL `NUMERIC(4)` data type definition follows the SQL standard, that designation is used for the target data type.

The data type precision (or size) and scale usually corresponds to the first and second parameters of the data type definition. However, the semantics for those parameters may also vary with the SQL implementation, requiring, for those cases, a specialized interpretation and conversion to an internal standard representation.

Due to the prioritization of the database content information over the database structural metadata, the data type conversion does not ensure that the target type will be the same as the original type, but rather a data type broad enough for all values that can be represented by the original data type, without any data losses (i.e. the target data type domain contains original data type domain). An example of this could be the conversion of a data type `VARCHAR(500)` (capable of holding a variable length sequence of up to 500 characters) to an hypothetical DBMS in which the maximum number of characters supported by the `VARCHAR` data type is 200. In this case, the module would choose a `TEXT` data type (provided that it had the capacity to hold 500 or more characters), ensuring that all character sequences could be represented by the target data type without any information loss.

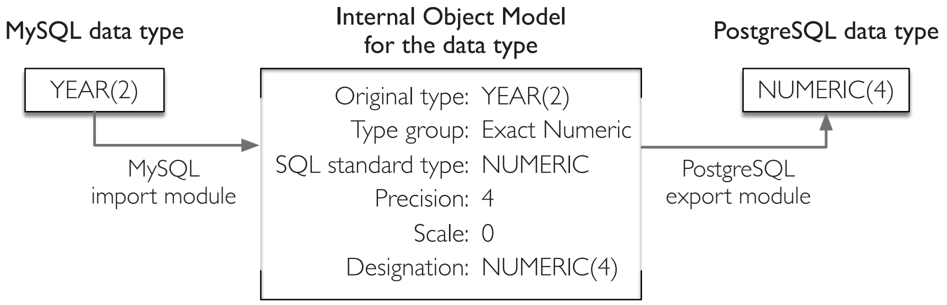


Figure 3: Conversion of MySQL YEAR data type to PostgreSQL

```
Database
├─Users
├─User roles
├─User privileges
└─Schemas
  ├─Tables
  │ ├─Columns
  │ ├─Primary key
  │ ├─Candidate keys
  │ ├─Foreign keys
  │ ├─Constraints
  │ └─Triggers
  ├─Views
  │ └─Columns
  ├─Routines
  └─User defined types
```

Figure 4: Database structure as an hierarchy.

The modules may also opt for using a different data type instead of a deprecated one. In some cases, the data type is changed to an alias data type with the exact same functionality, such as the `NUMERIC` and `DECIMAL` types on MySQL and Microsoft SQL Server.

During the conversion, data types and value changes are registered in a report file for manual analysis. This file may also be used as additional documentation for the generated SIARD file.

Some optimizations are also carried out by specific modules. One of those optimizations is implemented in all DBMS export modules and postpones adding database behavioural information until after the database content information is completely converted. If the behavioural information was to be added before the database content conversion, all inserted table records would be subject to a series of validations, such as primary key uniqueness or foreign keys presence, upon being inserted in the target database. Postponing the addition of these elements executes those validations only once, thus reducing the time needed to insert a table record in the target database. Also it allows to migrate the database even if constraints fail.

```
Database
├─Users
├─User roles
├─User privileges
├─Tables
│ ├─Columns
│ ├─Primary key
│ ├─Candidate keys
│ ├─Foreign keys
│ ├─Constraints
│ └─Triggers
├─Views
│ └─Columns
├─Routines
└─User defined types
```

Figure 5: Schema-less database structure as an hierarchy.

When converting the database contents, a flexible internal model must be used to represent different kinds of information and avoid data losses. The import module should select the most adequate model to be used for each record during a conversion.

Some values obtained from the database may not be in a standard format and must be converted to the standard format. A common example of such values are the `DATE`, `DATETIME`, `TIMESTAMP` and other date or time data types, because the format used internally by DBPTK is the ISO standard for representation of dates and times (ISO 8601)[10] and some dates are not provided in this format. Specific examples include the `YEAR` MySQL data type that must be converted to a numeric value in the range 1970 to 2069, inclusive.

3.1 Evaluating the Conversion

To ascertain the quality of the conversions made using DBPTK, a testing system was developed. The system was named *roundtrip testing* and is used to check if converting a database to a different Database Management System or format and then converting it back to the original Database Management System or format results in any data changes or losses.

[7] MySQL `YEAR` data type documentation available at http://dev.mysql.com/doc/refman/5.7/en/year.html

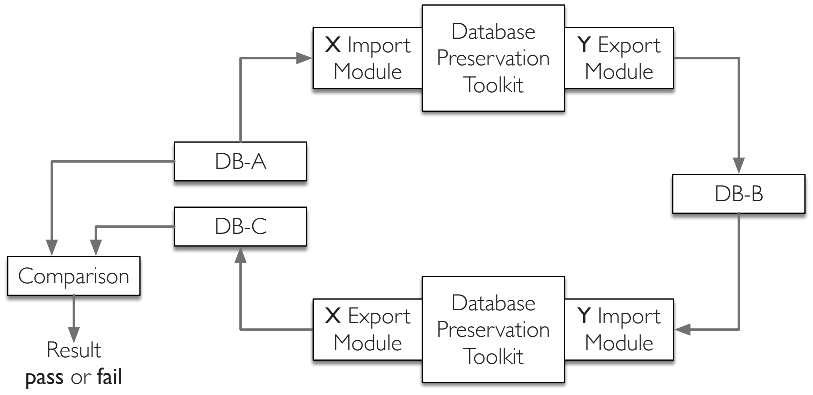


Figure 6: The *roundtrip test*

The *roundtrip test* is described below and illustrated in figure 6.

1. Create a new database, *DB-A*, in a Database Management System or format, *DBMS X*, with the intended test data (e.g. a table containing LOBs);

2. Use DBPTK to convert *DB-A* to a different Database Management System or format, *DBMS Y*, creating database *DB-B*;

3. Use DBPTK to convert *DB-B* back to *DBMS X*, creating database *DB-C*;

4. Compare *DB-A* and *DB-C*. The test passes if the database content information is unchanged.

Using this method, four modules are tested (two import modules and two export modules). The *roundtrip test* fails if any database content information is changed or lost during the conversions.

It is noteworthy that the comparison step may still consider the conversion to have succeeded when some database structure and behavioural information was lost or changed. This tolerance exists to accommodate for aliased data types and any incompatibilities between Database Management Systems or formats. An example of this is the `YEAR(2)` data type from MySQL, which is changed to `NUMERIC(4)` when converting to PostgreSQL (see figure 3) and would be created as `NUMERIC(4)` when converting the database back to MySQL.

4. DATABASE VISUALIZATION TOOLKIT

The preservation of databases is only successful if there is a way to access the archived databases. To accomplish this, the DBVTK is being developed, allowing archivists and consumers to preview and explore preserved databases in an intuitive interface.

The DBVTK is a scalable web-service that is able to serve multiple archived databases. It displays database description information, structural information, behaviour information and content, providing the ability to search and filter records from a single table or a whole database. Advanced search functionality allows filtering records using multiple search criteria and advanced data searches, such as searching date and time ranges. The DBVTK is optimized to provide almost instantaneous responses to searches on millions of records. Search results can then be exported to formats such as PDF (Portable Document Format) and CSV (Comma Separated Values).

When searching tables containing primary and foreign keys, it is often useful to be able to follow these relations and see the related records from the other table. This functionality in the DBVTK is triggered by clicking a cell containing a primary or foreign key, which will show the records from the other table related to the key. The database structural and description information can also be used to understand these relations.

The DBVTK can integrate with an external authentication and authorization system, providing the means to identify the users and verify their permissions to access each database.

After logging in, users will be able to see the list of databases they can access. By clicking one of the databases the user is shown some database metadata, such as the database name, description or data owner; from there the user can begin searching and exploring the database.

The DBVTK is not backed by a relational DBMS due to scalability and performance issues, instead the Apache Solr[8] platform is being used to store preserved database records. Apache Solr is an open source enterprise search platform. It was chosen for its versatility, scalability, and ability to provide almost instantaneous responses to searching and filtering queries on millions of records.

In order to provide access to preserved databases, the DBVTK requires the database to be loaded into Solr. This is achieved using DBPTK with a Solr export module. This module grants DBPTK the ability to add a SIARD database to a Solr platform such as the one used by the DBVTK (see the top part of figure 7).

As consumers use web-browsers to access the DBVTK web interface and explore databases, the back-end server application retrieves the database records and sends them to the web interface, which shows the records to the consumer (see the bottom part of figure 7).

[8] Apache Solr is available at http://lucene.apache.org/solr/

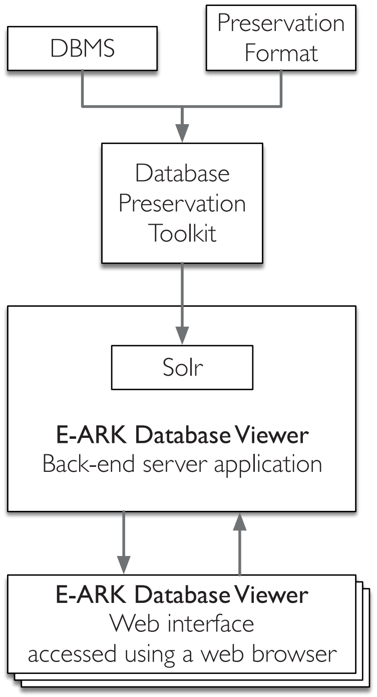

DBMS
Preservation Format
Database Preservation Toolkit
Solr
E-ARK Database Viewer
Back-end server application
E-ARK Database Viewer
Web interface
accessed using a web browser

Figure 7: Application architecture overview

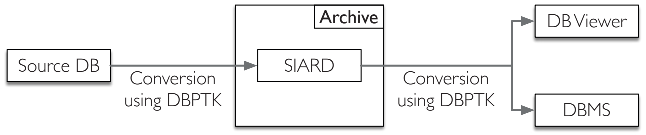


Figure 8: Usage scenario for an archive

5. USAGE SCENARIOS

The Database Preservation Toolkit and the Database Visualization Toolkit can be used in multiple cases to achieve different goals in an archive context.

1. Producer using DBPTK during the pre-ingest phase

During the pre-ingest phase, the producer can use DBPTK to convert a database to SIARD 2. After adding some documentation and other information, the producer can deliver the database to the archive. Such procedure is depicted in the left part of figure 8 and the following usage scenarios correspond to the right part of the same figure.

2. Consumer using DBVTK to explore a database

The archivist grants the consumer access to a database, and after logging in to the DBVTK web interface, the consumer is able to search and filter database records at will.

3. Consumer using DBVTK to explore a database prepared by an expert archivist (add views)

To serve a database with a specific view requested by a consumer, an archivist can use DBPTK to convert the SIARD database to a supported DBMS. The archivist can then use the DBMS to create the requested views, create a new SIARD using DBPTK. The new SIARD file can be exported to the DBVTK.

After being given access to the database, the consumer can access it using the DBVTK web interface to explore and search the records from the views.

4. Consumer using DBVTK to explore a database prepared by an expert archivist (serve only specific views)

An alternative to the previous method can be used when the archivist only wants to make part of the database information available to the consumer. By providing some options when creating the new SIARD file on DBPTK, the archivist may create a SIARD file containing a subset of the tables and views present in the source database.

Even after obtaining access to the new database, the consumer will only be able to access information present in the tables and views that were exported to SIARD. This particularly useful to restrict or completely block access to sensitive database information.

5. Researcher performing complex queries and analysis

A researcher may initially act as a consumer, requesting access and exploring databases until a suitable database is found. At that point, the researcher could obtain the suitable database in SIARD format, use DBPTK to convert it to a supported DBMS, and finally use the DBMS functionality to research the data. This allows a researcher to use Data Mining and OLAP techniques to research archived databases.

6. E-ARK PILOTS

The DBPTK is being piloted in the context of the E-ARK project by the Danish National Archives, the National Archives of Estonia and the National Archives of Hungary. [2]

The Danish National Archives pilot goal is to make four successful data extractions from live authentic databases into the SIARD 2.0 format:

- Extract records from Microsoft SQL Server database bigger than 100 GB (with a minimum success rate of 90%);

- Extract records from a large database containing documents;

- Extract records from Ms SQL database containing 50-60 tables and about 90.000 records (with a minimum success rate of 90%);

- Extract records from Microsoft SQL Server database containing about 5 million records.

One of the National Archives of Hungary goals is to convert an Oracle database to and from a preservation format, and accessing it using the DBVTK. This database is not normalized and contains more than 300.000 cases of the Hungarian Prosecution Office. The archives also aim to migrate two or more databases with different characteristics and containing both restricted and open content.

The National Archives of Estonia pilot aims to migrate a database with a complex structure and around 200.000 records.

The pilots demonstrate the potential benefits of these tools and how they can be used for easy and efficient access to archived records.

7. CONCLUSIONS AND FUTURE WORK

The Database Preservation Toolkit aims to support the migration of databases to the SIARD database preservation format and back to a DBMS. The SIARD format retains the database significant properties and its data can be validated using XML Schema Definition. Furthermore, by prioritizing the conversion, DBPTK ensures that no database content information is lost, and attempts to map the database structural and behavioural information to standard SQL, whilst keeping the original information as documentation. The software was made flexible to support different Database Management Systems and formats, including their specific features and optimizations. Moreover, DBPTK performs on low computing hardware requirements, even when converting databases containing millions of records.

The Database Visualization Toolkit aims to provide access to preserved databases, and achieves this by providing a fast and intuitive interface in which consumers can search and explore the preserved databases.

Both tools will be validated by the E-ARK pilots, by the European national Archives, ensuring that they are qualified to be used with real-world databases in a real archive environment.

Future work includes continuing the development of both tools, using means like the *roundtrip* tests and feedback from the E-ARK pilots to ensure top software quality and stability.

8. ACKNOWLEDGMENTS

This work is supported by the European Commission under FP7 CIP PSP grant agreement number 620998 – E-ARK.

9. REFERENCES

[1] ADDML - Archival Data Description Markup Language 8.2. Standard, Norwegian National Archives, Mar. 2011.

[2] I. Alföldi and I. Réthy. D2.3 - detailed pilots specification. Public deliverable, E-ARK, Mar. 2016.

[3] H. Bruggisser, G. Büchler, A. Dubois, M. Kaiser, L. Kansy, M. Lischer, C. Röthlisberger-Jourdan, H. Thomas, and A. Voss. eCH-0165 SIARD Format Specification. Standard, Swiss Federal Archives, Berne, Switzerland, 2013.

[4] E. F. Codd. A relational model of data for large shared data banks. *Commun. ACM*, 13(6):377–387, June 1970.

[5] T. M. Connolly and C. E. Begg. *Database Systems: A Practical Approach to Design, Implementation and Management (4th Edition)*. Pearson Addison Wesley, 2004.

[6] L. Faria, A. B. Nielsen, C. Röthlisberger-Jourdan, H. Thomas, and A. Voss. eCH-0165 SIARD Format Specification 2.0 (draft). Standard, Swiss Federal Archives, Berne, Switzerland, Apr. 2016.

[7] H. Heslop, D. Simon, and A. Wilson. *An Approach to the Preservation of Digital Records*. National Archives of Australia, Canberra, 2002.

[8] ISO 10646:2012, Information technology – Universal Coded Character Set (UCS). Standard, International Organization for Standardization, June 2012.

[9] ISO 19503:2016, Information technology – XML Metadata Interchange (XMI). Standard, International Organization for Standardization, Jan. 2016.

[10] ISO 8601:2004, Data elements and interchange formats – Information interchange – Representation of dates and times. Standard, International Organization for Standardization, 2004.

[11] ISO 9075:2008, Information technology – Database languages – SQL. Standard, International Organization for Standardization, 2008.

[12] M. Jacinto, G. Librelotto, J. C. Ramalho, and P. R. Henriques. Bidirectional conversion between XML documents and relational databases. In *The 7th International Conference on Computer Supported Cooperative Work in Design*. COPPE/UFRJ, 2002.

[13] J. C. Ramalho, M. Ferreira, L. Faria, and R. Castro. Relational database preservation through xml modelling. *International Workshop on Markup of Overlapping Structures (Extreme Markup 2007)*, 2007.

[14] RFC 1738, Uniform Resource Locators (URL). Standard, Internet Engineering Task Force (IETF), 1994.

[15] .ZIP File Format Specification, version 6.3.3. Standard, PKWARE Inc., Sept. 2012.

Exploring Friedrich Kittler’s Digital Legacy on Different Levels: Tools to Equip the Future Archivist

Jürgen Enge
University of Art and Design (FHNW) Basel
Freilager-Platz 1
4023 Basel, Switzerland
+41 61 228 41 03
juergen.enge@fhnw.ch

Heinz Werner Kramski
Deutsches Literaturarchiv Marbach
Schillerhöhe 8–10
71672 Marbach, Germany
+49 7144 848 140
kramski@dla-marbach.de

ABSTRACT

Based on the example of Friedrich Kittler’s digital papers at the Deutsches Literaturarchiv Marbach (DLA), this paper explores digital estates and their challenges on different logical levels within the pre-archival analysis, documentation and indexing process. As opposed to long-term digital preservation procedures, which are set about afterwards when relevant digital objects have already been identified, this process starts shortly after physical material (computers, hard drives, disks…) is delivered to the archive and has been ingested and safeguarded into volume image files. In this situation, it is important to get an overview of the “current state”: Which data was delivered (amount, formats, duplicates, versions)? What is the legal status of the stored data? Which digital objects are relevant and should be accessible for which types of users/researchers etc.? What kind of contextual knowledge needs to be preserved for the future? In order to address these questions and to assign meaning to both technological and documentation needs, the digital analysis tool “Indexer”[1] was developed [3]. It combines automated, information retrieval routines with human interaction features, thereby completing the necessary toolset for processing unstructured digital estates. It turns out however, that intellectual work and deep knowledge of the collection context still play an important role and must work hand in hand with the new automation efforts.

Keywords

Digital estates; digital papers; personal digital archives; digital analysis; file format identification; pre-archival process; semi-automated indexing; appraisal.

1. INTRODUCTION

The collections of the German Literature Archive (Deutsches Literaturarchiv – DLA) bring together and preserve valuable sources of literary and intellectual history from 1750 to the present day. Around 1,400 conventional papers and collections of authors and scholars, archives of literary publishers and about one million library volumes still make up the bulk of the collections. With the emergence of text processing and computer-assisted work for writers, authors and publishers, digital documents surely belong more and more to the field of collection of literary life and German-language contemporary literature.

With regard to digital unica – that usually remain unpublished and restricted to a single data carrier – a memory institution bears extraordinary responsibility for their long-term preservation, since per se no cooperative or redundant collection and indexing can be undertaken.

When DLA first began processing the digital estate of Thomas Strittmatter (1961–1995), it was one of the first memory institutions in the German-speaking world that needed to develop a workflow for digital unica [14, 77; 13]. Since then, 281 data carriers (almost exclusively 3.5"- and 5.25" floppy disks) from approximately 35 collections were saved, and roughly 26,700 files converted into stable formats.

With the exception of Strittmatter’s Atari and F.C. Delius’ Macintosh, only data carriers were acquired during this phase, but no complete computer environments. Often, disks were discovered incidentally while examining the conventional material, rather than deliberately acquired. Our priority was to conserve the texts as objects of information independent of their respective carriers. The two PCs were displayed in the Museum of Modern Literature (Literaturmuseum der Moderne – LiMo), but only as museum exhibition pieces, not as functional working environments in the sense of [5].

The digital estate of Friedrich Kittler (1943–2011), which was acquired in spring 2012 without any technical pre-custodial preparations, goes beyond the scope of previous procedures, both quantitatively and qualitatively. Thus, it became necessary to explore new options: Digital analysis tools and automated work routines have been brought into focus, in order to make the yet-unknown amounts of data manageable.

Friedrich Kittler was one of the most famous and important German media theorists and literary scholar. His impact on humanities in general and media studies in particular is of growing interest due to technological and methodological reasons. Since Kittler’s media archeological merits have derived to a great extent from his practical experiences in programming, it seems comprehensible that his intellectual legacy can only be understood and/or reconstructed by accessing both his theoretical work (books, articles, documented presentations) and his digital programming experiments. Whereas parts of the first were mostly published during his lifetime, the latter is basically hidden on nine hard drives, 104 optical disks, 648 floppies, etc. – hereinafter referred to as “Kittler’s digital estate”. Both are supposed to be (re-)edited in the now compiled Kittler edition.

Kittler bequeathed collected source codes as well as modifications of his own software and hardware configurations. Among the rather “idiosyncratically” [4, 14] structured data are thus “handwritten“ codes, like Kittlers 15 years spanning computer-based study of Markov chains, which “one might say, [forestall ...] Digital Humanities, since they constitute computer-based text analysis” [4, 12].

The wish to encounter scholarly pieces in their original, immediate environment and folder structure of Kittler’s personal computing working place made it necessary to show

[1] In jest we call the Indexer “Ironmaiden”: “Intelligent Read-Only Media Identification Engine” or “Intelligent Recursive Online Metadata and Indexing Engine” but the official name simply is “Indexer”.

utmost restraint in excluding data from future access. Even more so since Kittler routinely worked with root privileges, thus having and using writing authorizations everywhere. In other words, it was very important to find ways to make Kittler's source codes accessible – especially within their immediate local context at the original file location and position in the system directories. Therefore, all files from hard drives, most of the readable disks and about all optical media were examined. Only obvious "mass products" (such as CD-ROMs attached to the German computer magazine "c't") were only registered, but not copied.

Whereas from a technical point of view the heterogeneity of different file formats and the sheer mass of 1.7 million files were demanding, regarding semantic challenges it soon became clear that human interaction and decision making was indispensable. At the same time even these intellectual decisions had to be formulated in a rather formal way so that they could be applied to whole groups of (technically) similar data. Hundreds or thousands of files were just too much to analyze manually and the risk of publishing semantically restricted files was just too big.

2. IDENTIFYING AND DOCUMENTING

Since technological and semantic challenges of Kittler's digital estate did increase the documentation needs, implicit information had to be made explicit. Hidden knowledge had to be documented and assigned to its host components for enabling future investigations. As opposed to approaches which focus primarily on the content part of the digital estates and/or the raw files, the pre-archival indexing and appraisal processes meant in our case adding and keeping contextual information, too. Contextual information might be attached to the physical/hardware carrier (traces of handling) or conventions in naming or storing information at dedicated places, so careful documentation is recommended. Keeping track of this information supports later access regulations.

In his presentation, Christopher Lee 2012 defines eight "levels of representation" of digital resources [7, 7]. In contrast, we propose introducing an additional "level -1":

- Hardware (primarily as a museum object).

The sequence of our six levels is roughly related to the order of treatment. Combining a rather documentarian approach with institutional and operational needs in the pre-archival indexing process, we suggest furthermore at least five chunks of information entities:

- hard disks and data carriers (in terms of physical computing or storage media)
- (raw) disk images, which provide an important archival backup copy
- filesystems, indicating information about the used operating systems
- raw files, which contain the content/data
- context(ual) information, which is subsequently generated (in terms of virtual layer).

The following considerations start with the rather documentarian part which focuses on the first three levels: hardware, hard disk and data carrier, and image backup.

3. HARDWARE

The relevance of the hardware level again becomes apparent when considering the case of Kittler's estate: During April 2012, the DLA first received two older tower PCs from Kittler's estate, both of which had not been used for some time (his current PC was initially kept in Berlin, as a hardware reference for Kittler programs, and was at later date forwarded along with additional old laptops).

At first, from the perspective of conventional preservation raises the issue of cleaning the soiled and dusty hardware components. Due to the danger of carrying mold spores into the magazines, it was decided to remove loose dust, but to keep attached traces of grime and liquids as authentic signs of usage. For a reset button strewn with pen and pencil marks is a testimony of how often its adventurous user had to irregularly reboot his computer. Even after a complete migration and emulation of all digital objects, the hardware retains the nimbus of an original and potential exhibit.

During this early phase, it has proven valuable to decide on distinct (though not always chronologically correct) labels for the computers ("PC1", "PC2") and to keep a dossier with many photographs from the very beginning, in order to document and keep track of the growing amount of hardware information.

PC1 was brought to the archive without a hard drive, was non-functional and so was documented via visual examination only. With the help of live boot media (for example Xubuntu 8.10, which had to be used due to the limited RAM equipment) and tools like "lshw", "lspci", "lsusb", "hwinfo", "inxi" etc., the hardware configuration of PC2 and later, functional computers was analyzed.

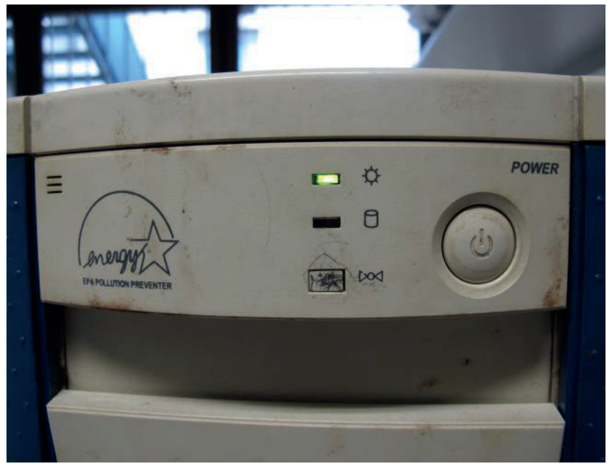


Figure 1. One of Kittler's old PCs (Pentium III, ca. 2000) showing heavy signs of usage on the reset button.

The inspection and analysis of the hardware required substantial employment of personnel, as well as profound IT knowhow, preferably with Linux distributions and hardware components of the period of use (such as SCSI hard drives and controllers). On the other hand, standardized live media and hardware diagnosis tools are available, which allow for a precise and fast overview. Apart from purely technical work, information about the usage context has to be collected, as this may influence the prioritization of tasks. For example, it became necessary to contact Kittler's former colleagues to learn his login password.

4. HARD DISKS AND DATA CARRIER

Very often data carriers are physically contextualized by the technological context in which they occur: a build-in hard drive fulfills different functions in most of the cases than a portable one. One might also differentiate semantically between a rather active usage of data carriers, which are continuously in use and thus integral part of the working process, and passive usage, in which data carriers are accessed only temporarily. Passive data carriers instead are often used for transporting data through time and place; they contain data which the owner kept with him/her for presentation or backing-up reasons, which might indicate a certain kind of relevance.

Since Kittler was a heavy smoker and a lot of dust settled down on data carriers stored under non-optimal conditions over the years, all volumes first entered the conservation and restoration team of DLA, which subjected the storage media to professional cleaning.

Before any further processing could be made, it had to be ensured that the write-protection of floppy disks was active. Because of the wide range of filesystems used on disks (including many "ext2" formatted ones), all reading operations have been carried out on Linux.

In a first reading step, all floppy disks were processed by a long command line which recorded – besides other technical metadata – the filesystem type and the change time of the most recent file contained on disk. This date was then temporarily attached to each volume by sticky notes and allowed manual re-consolidation of scattered disks to a joined set, for example a particular backup. This formation of groups could usually be confirmed by the disks' look and feel (make, labeling, signs of usage).

The cleaning and sorting was followed by a carefully designed labeling process, where internal identifiers were assigned to all hard disks and removable media.

- The acquired hard drives were distinctly labeled "hd01", "hd02" etc., which is to be understood as a numerus currens without chronological significance. A hierarchical attribution of internal hard disks to computers was not possible, since they were often either installed and functional, installed but not connected or completely separated with no way of determining which PC they belonged to.
- The naming of the contained partitions was largely based on another pattern, independent of the naming conventions of the running operating system. Other names for data carriers were defined as follows:
- fd001 etc.: floppy disks, disks
- od001 etc.: optical disks, CD-ROM, CD-R, CD-RW, DVD etc.
- xd001 etc.: external files: File collections on other external data carriers, e.g. on USB hard drives of the DLA.

The labels were written with a permanent marker on labeling boxes on cable ties, or on (mostly the backside) of the carriers themselves. For the labeling of black floppy disks, using "Posca" markers with water-solvable white pigment ink, the kind of which is also used by conservators, has proven successful.

These labels also served to create file names for sector images and listings and simplified the administration in internal lists that could later be imported into the Indexer. However, these labels are not identical to the archive's accession numbers, since those had not yet been assigned at that point.

Similar to hardware, inspecting, analyzing and possibly consolidating the data carriers required both substantial employment of staff and profound IT knowhow. However, via scripts and standard Linux tools ("mount", "ls" etc.) the analytical steps for disks can be conveniently automated. In Kittler's case, who archived numerous self-made copies of MS-DOS programs and operating systems on disks, knowledge of 1990s software is helpful for identifying and classifying these disks. Susanne Holl has shown that the frequency and occurrence of specific files on active and passive data carriers can reveal interesting information regarding relevance: "it is an interesting piece of information," she states, "that machine.txt was saved 22 times, itinerating through all hardware upgrades, from hard drive to floppy to hard drive to optical disk to hard drive" [4, 8].

Furthermore, close cooperation with one of his colleagues has been invaluable because she could identify many data carriers as Kittler's "writings" in the narrow sense of the word, which influenced the chronological order of further steps.

5. IMAGE BACKUP

Although DLA only began in 2014 (with the acquisition of Kittler's most recent PC) to actively use tools from the BitCurator distribution, almost from the beginning in 2003 it followed a strategy highly recommended by the BitCurator project: to conservate media volumes as a one-to-one copy into sector images, the "cornerstone of many forensics methods" [8, 27]. Recovery and analysis of deleted files is not part of DLA's standard workflow, but based on these images, it would at least be possible in cases of special need.

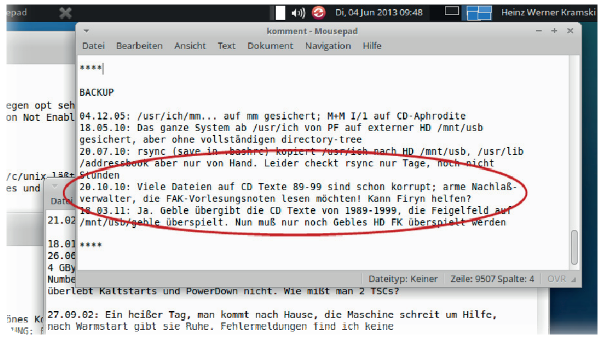
Figure 2. Running BitCurator (live medium) on Kittler's last PC (Intel Core i7-2600K, 2011).

In general, sector images are most qualified to preserve technical metadata of filesystems (update time stamps, user information etc.). Moreover, they can be directly integrated as logical drives (read-only) in virtual machines or emulators (see Figure 7).

Sector-wise copying of floppy disks could not be carried out with the previously used, custom-made windows tool "FloppImg" [13], because of the large amount of ext2 and other filesystems not mountable on Windows. A Linux script was used instead which calls the tool "ddrescue" and hence works well with deficient media.

244 disks out of a total of 648 were initially not considered during this work step, because they were obviously industrially produced installation disks for operating systems, drivers or application programs (MS-DOS, Windows 3.x, SCO Unix, Gentoo Linux) or 1:1 copies of the same. Their backup into the DLA software archive, which is established independently of Kittler and could be relevant to future emulations, is still pending. Whether these data carriers can be counted among Kittler's digital estate in the narrow sense, is open to debate. (When installed on his hard drives and theoretically executable, they certainly do, as they form his working environment.) But when in doubt, disks labeled either by handwriting or by typewriter were considered relevant and thus copied. Some disks were simple empty and not in use. However, disks that were apparently empty, but had handwritten labels were examined more closely using "Kryoflux". Out of 404 interesting candidates, it was in 119 cases not possible to create mount- and usable sector images. Therefore, the failure rate of Kittler's disks (the oldest ones date from 1987) amounts to 29.5%

CD-Rs instead were converted into .iso-files by the c't tool "h2cdimage" which creates partially usable images from deficient volumes like ddrescue [2]. In contrast to common copy programs it will not continue reading in deficient sectors without any further progress, so that the drive will not degrade from continued reading attempts.



```
****

BACKUP

04.12.05: /usr/ich/mm... auf mm gesichert; M+M I/1 auf CD-Aphrodite
18.05.10: Das ganze System ab /usr/ich von PF auf externer HD /mnt/usb
gesichert, aber ohne vollständigen directory-tree
20.07.10: rsync (save in .bashrc) kopiert /usr/ich nach HD /mnt/usb, /usr/lib
/addressbook aber nur von Hand. Leider checkt rsync nur Tage, noch nicht
Stunden
20.10.10: Viele Dateien auf CD Texte 89-99 sind schon korrupt; arme Nachlaß-
verwalter, die FAK-Vorlesungsnoten lesen möchten! Kann Firyn helfen?
18.03.11: Ja. Geble übergibt die CD Texte von 1989-1999, die Feigelfeld auf
/mnt/usb/geble überspielt. Nun muß nur noch Gebles HD FK überspielt werden

****
```

```
überlebt Kaltstarts und PowerDown nicht. Wie mißt man 2 TSCs?

27.09.02: Ein heißer Tag, man kommt nach Hause, die Maschine schreit um Hilfe,
nach Warmstart gibt sie Ruhe. Fehlermeldungen find ich keine
```

Figure 3. "Arme Nachlaßverwalter" (Poor curators)...

Regarding a central CD-R, Kittler says in the file "komment" (a kind of technical diary) "20.10.10: Many files on CD Texts 89–99 are already corrupt; poor curators who want to read FAK lecture notes!" [6]. It is remarkable to be addressed from the past in such a way. It is also remarkable how of all things, it was Kittler's beloved c't that helped save an unexpectedly high portion of backup CD-Rs that he had already dismissed as "broken" during an interview [10].

Out of 104 optical data carriers, 82 were temporarily ruled out as mass-produced ware and installation media. Out of the remaining 22 self-burned CD-Rs, only three could not be flawlessly copied. However, it was possible to later mount them.

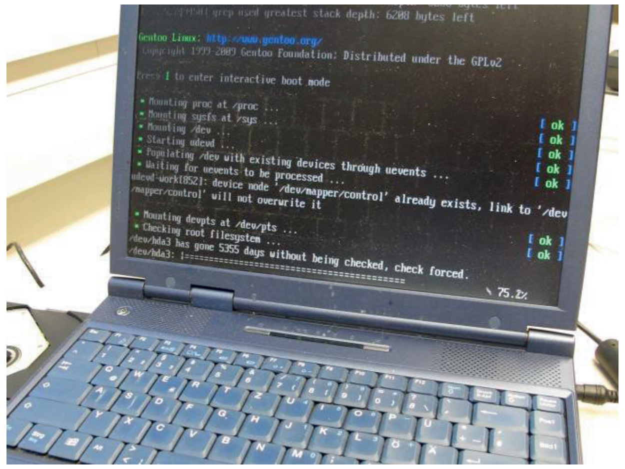
Hard drive partitions were at first also created using Linux scripts and ddrescue. From 2014 onwards, "Guymager" in "dd" mode (without file splitting) was used. Regrettably, there was an unreadable partition on a 2 GB SCSI drive.

Besides the principal difficulties of selecting relevant files for file format migration and for further editing, real technical problems arose in the attempt to store original files from hard disk partitions and optical volumes on the standard file servers of the DLA (as it was previously possible with the floppy disk inventory):

1. A digital estate is stored on the file server with an extensive path named after its holder with systematically labeled subfolders according to the processing state (see [12]). If original files to be stored have their own, deeply nested path hierarchy, the allowed path length of the operating systems involved might be exceeded.
2. Today's virus scanners often impede the copying of original files contaminated with old (MS-DOS) viruses.
3. DLA's standard file server does not support the original case-sensitive file names (e.g. Makefile vs. makefile) when serving Windows-based clients.
4. Reading errors often prevent file-by-file copying of original media.

It is possible to overcome all these limitations by mounting disk images, but then an appropriate presentation tool is needed. The Indexer therefore not only is required for full text indexing and MIME type analysis (see section 7), but also serves as a document server which preserves the authentic path information. However, the main motivation for developing and applying the Indexer remains the fact that 1,7 million files cannot be assessed by our colleagues in the archive without prior technical preparation, while at the same time, all technical measures must concentrate on a selection that can only be made through intellectual assessment. An implicit decision of relevancy, as it was possible in case of floppy disks, is bound to fail, when it comes to the enormous amounts of data contained by hard disks.

Although the primary reason for image copies are archival needs (backup, protection of the original source), they also offer a starting point for the indexing process, which can only start when an accessible filesystem is available.



```
grep used greatest stack depth: 6208 bytes left

Gentoo Linux: http://www.gentoo.org/
Copyright 1999-2009 Gentoo Foundation: Distributed under the GPLv2

Press I to enter interactive boot mode

 * Mounting proc at /proc ...                                          [ ok ]
 * Mounting sysfs at /sys ...                                          [ ok ]
 * Mounting /dev ...                                                   [ ok ]
 * Starting udevd ...                                                  [ ok ]
 * Populating /dev with existing devices through uevents ...           [ ok ]
 * Waiting for uevents to be processed ...
udevd-work[852]: device node '/dev/mapper/control' already exists, link to '/dev
/mapper/control' will not overwrite it

 * Mounting devpts at /dev/pts ...                                     [ ok ]
 * Checking root filesystem ...                                        [ ok ]
/dev/hda3 has gone 5355 days without being checked, check forced.
/dev/hda3: |===========================================
                                                                  \ 75.2%
```

Figure 4. Why you should do disk imaging before anything else: "/dev/hda3 has gone 5355 days without being checked, check forced".

For disk imaging there is very good tool support and long running copy or checksum jobs can easily be done on the side. Still, all steps have to be carefully monitored and documented, so IT knowhow is of advantage. However, as soon as data carrier identification has taken place and more detailed task schedules can be prepared, producing specific disk images can be delegated.

6. ANALYZING

Regarding the previously mentioned chunks of information, the analyzing part of the information retrieval starts at the filesystem level. It is followed by the raw files themselves and ends with observations regarding the contextual information. The result of the analysis of the filesystem are intentional statements because (at least parts of) the filesystem contain information about working process and conventions of the author: "We constantly seek not an artificially imposed classification by subject, but authentic pattern recognition of media in their archival order" [11, 112]. Kittler, for example, used several operating systems in parallel, including MS-DOS, SCO-Unix, early Windows versions and later primarily Gentoo Linux, which identify themselves due to their file structure. Furthermore and as already stated, he preferred working as "root" on Linux, bypassing administrative limitations normally applied to standard users. His standard working directory was not the commonly used subfolder of "/home", but "/usr/ich" instead. At first glance, Kittler seems to place himself on one level with system directories under "/usr" in the filesystem hierarchy. It is more likely, however, that he simply continued a convention of his earlier SCO Unix, which did indeed place user directories under "/usr". Still, the naming of his user account as "ich" (Me) certainly shows that he did not consider himself "one of several" users of his computers.

Inside his working directory a semantic order is largely missing, since he organized his files based on their character set: ASCII (",asc"), Latin9 (".lat"), UTF8 (".utf") [1, Min.: 13.50f]. Also, the usage of non-standard file extensions made an automated MIME type identification useful.

7. FILESYSTEM

Independent of Kittler's case, information of the filesystem comes in general close to classical cataloging information as far as author, title, date of creation, format etc. are recorded.

As preparation, the created sector images were made available to the Indexer VM via a read-only NFS share. There, they were mounted as loopback devices ("basepath" in table 1). To be able to use hundreds of these devices, a kernel parameter had to be raised. There was a highly specialized IRIX filesystem (XFS using "version 1" directories), for which current Linux systems no longer provide drivers. However, this could be mounted using a very old version of Ubuntu (4.10 with kernel 2.6.8) and copied on ext3, which, in this special case, provided the base for further steps.

From the documentation described in section 2.3, a list was loaded into the Indexer which assigned a unique session ID and a short description to every image (see Table 1).

For collecting and producing technical metadata, the Indexer first reads the ID of the archiving sequence (sessionid) specified on the command line for a particular image container and Indexer run. Then for each (recursively detected) filesystem object a distinct file identification number is generated (fileid), which refers to this specific indexing session. Another ID (parentid) identifies the folder, in which the directory entry is filed, and finally the file or folder name referred (name). The path of the directory entry is documented (path) as well as the basic type (filetype), for instances such as "file", "directory", "reference", the size of the file (filesize), and a checksum (sha256), which can be used for authenticity verification purposes.

Table 1. Session table of the Indexer (simplified excerpt).

sessionid	name	basepath	localpath	...
2001	hd01-p01	/Primaerbestand-mounted/ Kittler,_Friedrich_Adolf/ 0_Original-Disk/hd/hd01/p01	/u01/fk/hd/	
2002	hd01-p02	/Primaerbestand-mounted/ Kittler,_Friedrich_Adolf/ 0_Original-Disk/hd/hd01/p02	/u01/fk/hd/	
3001	od001	/Primaerbestand-mounted/ Kittler,_Friedrich_Adolf/ 0_Original-Disk/od/od001	/u01/fk/od/	
3002	od002	/Primaerbestand-mounted/ Kittler,_Friedrich_Adolf/ 0_Original-Disk/od/od002	/u01/fk/od/	
4001	fd001	/Primaerbestand-mounted/ Kittler,_Friedrich_Adolf/ 0_Original-Disk/fd/fd001	/u01/fk/fd/	
4002	fd002	/Primaerbestand-mounted/ Kittler,_Friedrich_Adolf/ 0_Original-Disk/fd/fd002	/u01/fk/fd/	

...	group	bestand	description	solrpath	
	hd	kittler	Partition 0,4 GB vfat, ca. 20040000, 1. Partition auf hd01 (IBM Deskstar, 32 GB, IDE) aus PC2	/solr/kittler	
	hd	kittler	Partition 15,7 GB ext3, ca. 20030000, 2. Partition auf hd01 (IBM Deskstar, 32 GB, IDE) aus PC2	/solr/kittler	
	od	kittler	CD-R iso9660, ca. 20010820	/solr/kittler	
	od	kittler	CD-R iso9660	/solr/kittler	
	fd	kittler	3,5" vfat, ca. 19900300	/solr/kittler	
	fd	kittler	3,5" vfat, ca. 19900300	/solr/kittler	

Later this is also double checked with entries of the National Software Reference Library (NSRL) of the American National Institute of Standards and Technology (NIST) in order to identify registered files of common software packages [9].

Furthermore, the date/time stamps when files were changed (filectime) or last accessed (fileatime) are of great importance. Care must be taken here to prevent unintentional modifications to the time attributes, so all containers strictly may not be mounted in write mode. Last but not least, all information of the Unix-call stat() ("stat") and the indexing time and date ("archivetime") are documented.

For storing this basic information and in preparation of the later full-text index, the Indexer maintains a directory of all filesystem objects and their technical metadata in a MySQL database. Metadata created during the information retrieval, as well as the information on the access path is stored beside the record. (The importance of the original path is emphasized by the implemented quotation routine, which displays an APA-like reference for citation). The naming convention of the session ID allows the administration of different filesystems/different estates or groups of objects. To uniquely refer to a single file a combination of sessionid and fileid is recommended.

During the first run, a copy of each file also is written into a balanced cache folder ("localpath" in table 1), so the image containers do not need to be present all the time. This also overcomes most of the limitations of common file servers outlined in section 5 and allows providing file links to the user without access to the archived sector images.

8. RAW FILES

Since the 'raw files' are supposed to contain the content of information itself, their analysis is of special importance. The iterative identification cascade of the Indexer analyzes the data step-by-step and optimizes the identification quality. Since every file identification tool has its own particular qualities and shortcomings, the Indexer combines different software tools. The list can also be changed, replaced or upgraded at any time. The varying results derive from different recognition algorithms and -databases within the single tools. Since contradictory statements can occur, the Indexer treats all results as equal, so that the user has to decide which information he or she would trust.

Among the mandatory tools the following software packages are of special importance: "Libmagic", which creates the initial list of files and tries to identify MIME type and encoding, and "gvfs-info", which has similar capabilities, but can sometimes deliver different results.

Highly recommended is furthermore "Apache Tika", which extracts not only the MIME type and encodings, but also the full text in case of texts. Extracted full texts are compressed with "gzip" to save cache space. "avconv/ffmpeg" is then used for extracting technical metadata from files, which "gvfs-info" has already identified as time based media (MIME type "video/*" or "audio/*"). "ImageMagick" is finally consulted for analyzing image- and PDF-data, of which it creates thumbnails. These thumbnails are used as preview images in the user interface.

In addition, "Detex" is useful for extracting the content (text) from TeX-files (MIME type "text/x-tex") by removing the TeX-commands. "Antiword" extracts full text from older Word-files (MIME type "text/application-msword), and "xscc.awk" extracts comments from the source code. The NSRL (locally imported into a Berkeley DB for performance reasons), which was already mentioned, is used for identifying software, which was not modified but only used by Kittler. The "md5sum" creates a checksum in one of the required formats, when matching against the NSRL is done.

The Indexer's core is a "SOLR" full text index. It collects the results of the iterative identification cascade in a separate, searchable index. This is mainly for performance reasons, but it also provides an autonomous subsystem, which is independent of the indexing and MySQL infrastructure. The full-text index itself is made accessible through a web-based user-interface, which enables search and information retrieval.

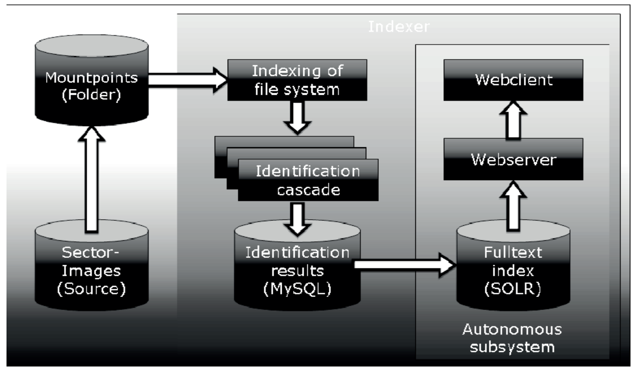
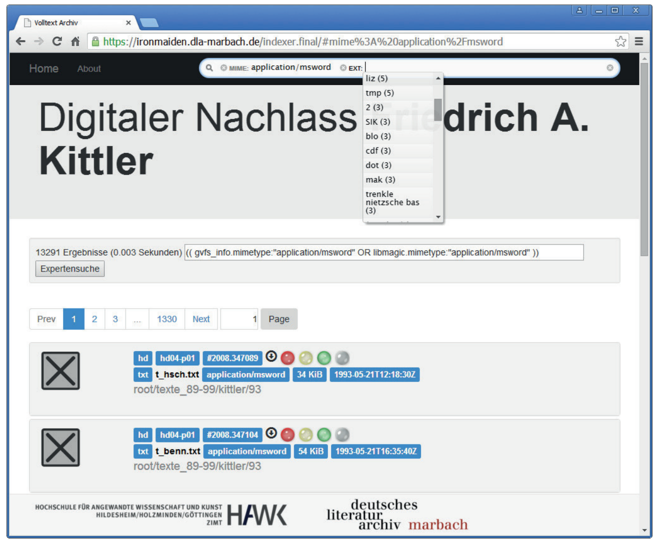


Figure 5. Indexer system architecture.

The simplified scheme above shows the overall system architecture of the indexer: Due to the large data volume, the Indexer runs were time-consuming and had to be gradually initiated and monitored. However, this effort is very much worthwhile: The knowledge gained through the automated MIME type analysis can hardly be overstated, since the estate is, from a traditional perspective, still unindexed. For example, a manual inspection might have classified word files with the extensions .doc, .DOC, .txt, .TXT, .dot, .DOT etc. as relevant for further investigation and possible migration of file formats. Unconventionally-labeled word files such as "*.vor" (presumably "Vorwort", preface) or "*.liz" might have escaped notice altogether.


Volltext Archiv
Home
About
HOCHSCHULE FÜR ANGEWANDTE WISSENSCHAFT UND KUNST HILDESHEIM/HOLZMINDEN/GÖTTINGEN ZIMT
HAWK
deutsches literatur archiv marbach

https://ironmaiden.dla-marbach.de/indexer.final/#mime%3A%20application%2Fmsword

MIME: application/msword EXT:

liz (5)
tmp (5)
2 (3)
SIK (3)
blo (3)
cdf (3)
dot (3)
mak (3)
trenkle nietzsche bas (3)

Digitaler Nachlass Friedrich A. Kittler

13291 Ergebnisse (0.003 Sekunden) ((gvfs_info.mimetype:"application/msword" OR libmagic.mimetype:"application/msword"))

Expertensuche

Prev 1 2 3 ... 1330 Next 1 Page

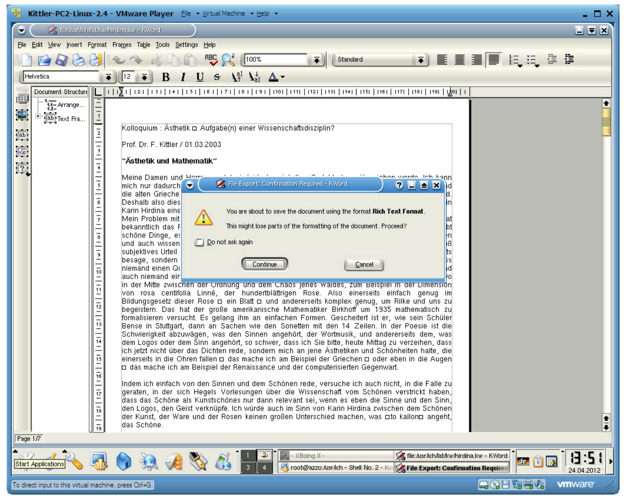
hd hd04-p01 #2008.347089
txt t_hsch.txt application/msword 34 KiB 1993-05-21T12:18:30Z
root/texte_89-99/kittler/93

hd hd04-p01 #2008.347104
txt t_benn.txt application/msword 54 KiB 1993-05-21T16:35:40Z
root/texte_89-99/kittler/93

Figure 6. Searching for unusual MS-Word file extensions.

It must be noted however, that DLA has currently just completed the bitstream preservation work and did not yet enter the stage of systematic file format migration. Besides MS-Word, Kittler mainly used Emacs for text editing, so in the areas of scientific papers and source code, his digital estate should not impose too much future problems.

One notable exception are KWord files (".kw") for which no known migration tools seem to exist – even the direct successor, KDE's "Calligra" suite is unable to import the older, proprietary (pre-2005) ".kw" format. In a singular, important case, a Kittler Linux machine was brought to life again as a virtual machine and allowed to save these documents as ".rtf" files for further processing. But in general, virtualization (or emulation) currently requires too many manual arrangements to be part of an efficient standard workflow and will be addressed in particular by the planned edition of Kittler's collected writings, in whose edition plan a part for his own software projects is explicitly included.


Kittler-PC2-Linux-2.4 - VMware Player
File
Virtual Machine
Help
file:/usr/ich/lat/kw/hirdina.kw - KWord
File Edit View Insert Format Frames Table Tools Settings Help
100%
Standard
Helvetica
12
Document Structure
Arrange...
Text Fra...
Page 1/7
Start Applications
- XBoing II -
root@azzo:/usr/ich - Shell No. 2 - Ko
file:/usr/ich/lat/kw/hirdina.kw - KWord
File Export: Confirmation Required
8:51
24.04.2012
vmware

Kolloquium : Ästhetik □ Aufgabe(n) einer Wissenschaftsdisziplin?

Prof. Dr. F. Kittler / 01.03.2003

"Ästhetik und Mathematik"

Meine Damen und Herr... mich nur dadurch ... die alten Grieche ... Deshalb also dies ... Karin Hirdina eine ... Mein Problem mit ... bekanntlich das F ... schöne Dinge, es ... und auch wissen ... subjektives Urteil ... besage, sondern ... niemand einen Qu ... auch niemand eir ... in der Mitte zwischen der Ordnung und dem Chaos jenes Waldes, zum Beispiel in der Dimension von rosa centifolia Linné, der hundertblättrigen Rose. Also einerseits einfach genug im Bildungsgesetz dieser Rose □ ein Blatt □ und andererseits komplex genug, um Rilke und uns zu begeistern. Das hat der große amerikanische Mathematiker Birkhoff um 1935 mathematisch zu formalisieren versucht. Es gelang ihm an einfachen Formen. Gescheitert ist er, wie sein Schüler Bense in Stuttgart, dann an Sachen wie den Sonetten mit den 14 Zeilen. In der Poesie ist die Schwierigkeit abzuwägen, was den Sinnen angehört, der Wortmusik, und andererseits dem, was dem Logos oder dem Sinn angehört, so schwer, dass ich Sie bitte, heute Mittag zu verzeihen, dass ich jetzt nicht über das Dichten rede, sondern mich an jene Ästhetiken und Schönheiten halte, die einerseits in die Ohren fallen □ das mache ich am Beispiel der Griechen □ oder eben in die Augen □ das mache ich am Beispiel der Renaissance und der computerisierten Gegenwart.

Indem ich einfach von den Sinnen und dem Schönen rede, versuche ich auch nicht, in die Falle zu geraten, in der sich Hegels Vorlesungen über die Wissenschaft vom Schönen verstrickt haben, dass das Schöne als Kunstschönes nur dann relevant sei, wenn es eben die Sinne und den Sinn, den Logos, den Geist verknüpfe. Ich würde auch im Sinn von Karin Hirdina zwischen dem Schönen der Kunst, der Ware und der Rosen keinen großen Unterschied machen, was □to kallon□ angeht, das Schöne.

File Export: Confirmation Required - KWord

You are about to save the document using the format **Rich Text Format**.

This might lose parts of the formatting of the document. Proceed?

☐ Do not ask again

Continue   Cancel

To direct input to this virtual machine, press Ctrl+G.

Figure 7. A Kittler VM running KWord version 1.2.1.

9. CONTEXTUAL INFORMATION

Beyond the technical analysis of data (indexing cascade), additional options for filtering are required. Since personal computers tent to contain private content (some may even be locked for 60 years by arrangement), information which touches third party personal rights (e.g. evaluation files), collected materials etc. withdrawing access rights from documents is essential. In case of DLA, suspending data is subject only to specific security measures and can only be imposed or removed by the identifiers and the heiress. Questionable content can be added with a disclaimer, which informs the user that the data can't be accessed due to further specified reasons. The instance or file can thus still be referred to in the pre-view. Via a self-explaining 'traffic light' system access-rights can be visualized and changed.

Table 2. Indexer access levels

🔓 Indexer record is unlocked (visible)

- 🔴 Show technical object ID only
- 🟡 Show metadata only
- 🟢 Show metadata and content,
show fulltext search results in multi-line context,
allow download (on campus)
- ⚪ Undefined, needs review

🔒 Indexer record is locked (invisible; visible only to administrators)

Whereas withdrawal of usage rights can only be triggered by defined users who obtain specific editor rights and/or authorized scholars, locking off specific files, all other rights can only be set by administrators.

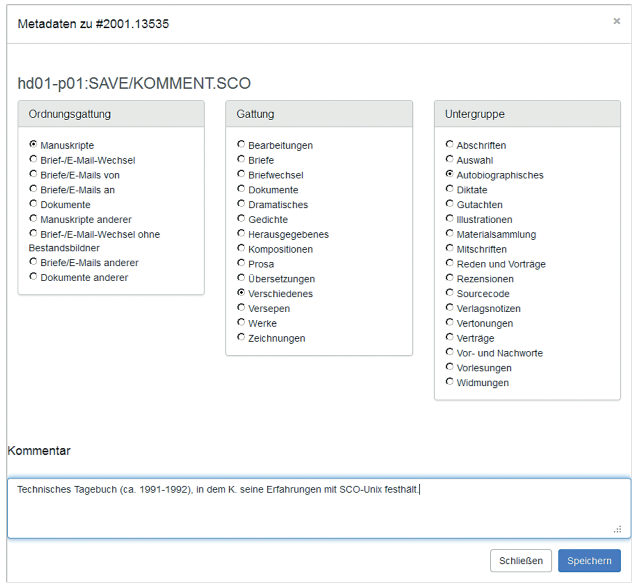
To execute mass classification which follows this scheme, formal rules have been created, which use server-side scripts. Among the applied routines are the following logical operations:

- Blur all thumbnails and set access level "Red" for all files having "mimetype:'image/jpeg'". This causes that all private photos get protected; however a great number of unproblematic images gets hidden as well. Another example of this type may be: Set access level "Yellow" for all files having "application/mbox" or "message/rfc822". This protects the content of all incoming or outgoing emails. These rules can easily be applied by some SQL statements as the MIME type is already known, so the degree of automatization is high.
- Set a specific access level for selected folders or file names (which are known) to be especially problematic ("Red" or "Yellow") or especially unproblematic ("Green"). This only works based on lists created manually by Kittler's widow, to whom the inventory is well known. It also works only because Kittler's use of folder and file names remained quite stable over the years and through different (backup) volumes. However, manual work involved in this step is high and the risk of missing some problematic files or folders cannot be eliminated.
- Set access level "Green" for all files found in the NSRL. This is easy to do, but unfortunately only covers the less interesting files. (At least, it reduces the amount of files to be processed further by roughly a third or 570,000.)

Setting and checking access levels is still ongoing work.

Another type of contextual Information, which follows the principle of metadata enrichment, is implemented for future use with a checkbox system. Scholars and/or editors can classify entries according to DLA's standard classification with respect to the content. Additional features like a discussion forum might be appropriate to add in the midterm.

Currently filtering options are primarily meant to support the preliminary classification process or to filter data which is not yet meant to enter the public sphere.



Metadaten zu #2001.13535 ×

hd01-p01:SAVE/KOMMENT.SCO

Ordnungsgattung

● Manuskripte
○ Brief-/E-Mail-Wechsel
○ Briefe/E-Mails von
○ Briefe/E-Mails an
○ Dokumente
○ Manuskripte anderer
○ Brief-/E-Mail-Wechsel ohne Bestandsbildner
○ Briefe/E-Mails anderer
○ Dokumente anderer

Gattung

○ Bearbeitungen
○ Briefe
○ Briefwechsel
○ Dokumente
○ Dramatisches
○ Gedichte
○ Herausgegebenes
○ Kompositionen
○ Prosa
○ Übersetzungen
● Verschiedenes
○ Versepen
○ Werke
○ Zeichnungen

Untergruppe

○ Abschriften
○ Auswahl
● Autobiographisches
○ Diktate
○ Gutachten
○ Illustrationen
○ Materialsammlung
○ Mitschriften
○ Reden und Vorträge
○ Rezensionen
○ Sourcecode
○ Verlagsnotizen
○ Vertonungen
○ Verträge
○ Vor- und Nachworte
○ Vorlesungen
○ Widmungen

Kommentar

Technisches Tagebuch (ca. 1991-1992), in dem K. seine Erfahrungen mit SCO-Unix festhält.

Schließen   Speichern

Figure 8. Indexer classification system.

10. CONCLUSION

As should be shown by the article, assigning meaning to digital information is indispensable while facing topics such as long-term access and sustained understanding, research data cycles the preservation needs and the mediation of contextual information over time. Whereas automatized indexing routines enable preserving content, a first result of human interaction is given by a number of grouping routines, which could be established in collaboration with selected archivists and editors. Relating technologically and semantically connected clusters of data with each other, as explained before, provides a good example how far technological skills and semantic knowledge can go hand in hand.

Choosing a less common way of argumentation, our survey tried to explain how far both sides can profit from each other. Whereas parts of the mentioned tasks may be conducted more and more often (and better) by digital tools such as the Indexer, others still require skilled archivists which are familiar with both worlds: the humanities as a field which enables identifying, assigning and documenting meaning in terms of culture, historical, or additional semantic values, and computer science, since technology and their identification get more and more complex.

This leads to at least two points: First, regarding current education and training facilities, a need to cover the cross mix of assigned competences becomes obvious. Especially in Europe, where digital realities in the heritage context have been neglected for too long, certain changes seem to be required. Existing education facilities need to be expanded and at the same time become more attractive to people from different fields of humanities as well as information science. At the same time an image change is required, which deconstructs the cliché of digital culture as nerdy and/or low culture.

The second aspect occurs by facing the big picture of current preservation approaches: Here it seems that (at least) two different types of interest motivate preservation actions today: a) the re-use of data and b) sustainability of authenticity. Whereas in the science sector a strong motivation for (scientific and/or economic) re-use can be observed, ensuring authenticity seems to be the primary aim within the cultural context of memory institutions. Both principles do not necessarily oppose each other. In practice, nevertheless, they can lead to the implementation of varying preservation strategies, parameters and solutions. One example can be found in comparing the way how significant properties or preservation priorities are defined. Archives such as the DLA are positioned at the vertex of these two lines: On the one hand, they are legally bound to preserve the authenticity in the sense of cultural identity. At the same time and at an increasing rate, they are subject to science and the standards of accessibility. However, this intermediate position makes archival involvements in digital preservation actions so interesting. Being routed in both spheres, interest groups of different areas can profit from each other. In this regard, the case of Friedrich Kittler can be seen as paradigmatic: his heritage in humanities will stay only partially comprehensible, without sufficient technical knowledge and vice versa.

11. ACKNOWLEDGEMNTS

Our thanks go to Susanne Holl, Kittler's widow, and to Tania Hron, his colleague, as well as all of his former colleagues. They have all been an invaluable help in technical and especially content-related questions and made us feel the importance of the work we did on Kittler's digital estate.

12. REFERENCES

[1] Berz, Peter and Feigelfeld, Paul. 2013. Source Code als Quelle. Aus der Arbeit mit Friedrich Kittlers Programmierwerk. TU-Lecture, (Berlin, Germany, Aug. 6 2013). URL=https://www.youtube.com/watch?v=kOjGcrj47rk.

[2] Bögeholz, Harald. 2005. Silberpuzzle. Daten von beschädigten CDs und DVDs retten. *c't – Magazin für Computertechnik*, 16 (2005), 78–83. URL=https://shop.heise.de/katalog/silberpuzzle.

[3] Enge, Jürgen, Kramski, Heinz Werner and Lurk, Tabea. 2014. Ordnungsstrukturen von der Floppy zur Festplatte. Zur Vereinnahmung komplexer digitaler Datensammlungen im Archivkontext. In *Beiträge des Workshops "Digitale Langzeitarchivierung" auf der Informatik 2013* (Koblenz, Germany, Sept. 20 2013), 3–13. URN=urn:nbn:de:0008-2014012419.

[4] Holl, Susanne. 2016. Friedrich Kittler's Digital Legacy. Part II. Forthcoming.

[5] Kirschenbaum, Matthew, Farr, Erika L., Kraus, Kari M., Nelson, Naomi, Peters, Catherine Stollar, Redwine, Gabriela. 2009. Digital Materiality: Preserving Access to Computers as Complete Environments. In *iPRES 2009. The Sixth International Conference on Preservation of Digital Objects* (San Francisco 2009), 105–112. URL=http://escholarship.org/uc/item/7d3465vg#page-1.

[6] Kittler, Friedrich. 2011. komment. Computer file. (#1001.10531, text/x-c, 2011-08-18T14:37:46Z,. In: Bestand A:Kittler/DLA Marbach. xd002:/kittler/info [xd, 352.4 KiB]). Internal URL=https://ironmaiden.dla-marbach.de/indexer.final/#id%3A%201001.10531.

[7] Lee, Christopher. 2012. Archival Application of Digital Forensics Methods for Authenticity, Description and Access Provision. (International Council on Archives Congress, Brisbane, Australia, August 20–24, 2012). URL=http://ils.unc.edu/callee/ica-2012-lee.pdf.

[8] Lee, Christopher, Woods, Kam, Kirschenbaum, Matthew and Chassanoff, Alexandra. 2013. From Bitstreams to Heritage. Putting Digital Forensics into Practice in Collecting Institutions. BitCurator Project. URL=http://www.bitcurator.net/wp-content/uploads/2013/11/bitstreams-to-heritage.pdf.

[9] NIST. 2016. Introduction to the NSRL. URL=http://www.nsrl.nist.gov/new.html.

[10] Rosenfelder, Andreas and Kittler, Friedrich. 2011. Wir haben nur uns selber, um daraus zu schöpfen (Interview). *Welt am Sonntag,* (Jan 30 2011). URL=http://www.welt.de/print/wams/kultur/article12385926/Wir-haben-nur-uns-selber-um-daraus-zu-schoepfen.html.

[11] Taylor, Hugh A. 1982/3. The Collective Memory. Archives and Libraries. In: *Archivaria 15*. URL= http://journals.sfu.ca/archivar/index.php/archivaria/article/view/10975/11908.

[12] von Bülow, Ulrich. 2003. Rice übt Computer, die Laune wird immer guter! Über das Erschließen digitaler Nachlässe (KOOP-LITERA Symposium, Mattersburg, Österreich, May 08–09, 2003). URL=http://www.onb.ac.at/koop-litera/termine/kooplitera2003/Buelow_2003.pdf.

[13] von Bülow, Ulrich, Kramski, Heinz Werner. 2011. Es füllt sich der Speicher mit köstlicher Habe. Erfahrungen mit digitalen Archivmaterialien im Deutschen Literaturarchiv Marbach. *Neues Erbe: Aspekte, Perspektiven und Konsequenzen der digitalen Überlieferung.* Karlsruhe, 141–162. DOI=http://dx.doi.org/10.5445/KSP/1000024230.

[14] Weisbrod, Dirk. 2015. *Die präkustodiale Intervention als Baustein der Langzeitarchivierung digitaler Schriftstellernachlässe*. Doctoral Thesis. Humboldt-Universität zu Berlin. URN=urn:nbn:de:kobv:11-100233595.

Persistent Web References – Best Practices and New Suggestions

Eld Zierau
Digital Preservation Dept.
The Royal Library of Denmark
P.O. Box 2149
1016 Copenhagen K - Denmark
+45 9132 4690
elzi@kb.dk

Caroline Nyvang
National Heritage Collection Dept.
The Royal Library of Denmark
P.O. Box 2149,
1016 Copenhagen K - Denmark
+45 9132 4919
cany@kb.dk

Thomas Hvid Kromann
National Heritage Collection Dept.
The Royal Library of Denmark
P.O. Box 2149
1016 Copenhagen K - Denmark
+45 9132 4422
thok@kb.dk

ABSTRACT

In this paper, we suggest adjustments to best practices for persistent web referencing; adjustments that aim at preservation and long time accessibility of web referenced resources in general, but with focus on web references in web archives.

Web referencing is highly relevant and crucial for various research fields, since an increasing number of references point to resources that only exist on the web. However, present practices using URL and date reference cannot be regarded as persistent due to the volatile nature of the Internet, - and present practices for references to web archives only refer to archive URLs which depends on the web archives access implementations.

A major part of the suggested adjustments is a new web reference standard for archived web references (called wPID), which is a supplement to the current practices. The purpose of the standard is to support general, global, sustainable, humanly readable and technology agnostic persistent web references that are not sufficiently covered by existing practices. Furthermore, it can support better practices for precise references in spite of the temporality issues for web material as well as issues related to closed web archives.

In order to explain needed change of practices based on the wPID, the paper includes a thorough description of the challenges in web references. This description is based on the perspectives from computer science, web collection and Digital Humanities.

Keywords

Persistent identification, Web references, Web Persistent Identifiers (wPID), Web elements, Digital Preservation.

1. INTRODUCTION

The main goal of this paper is to suggest needed changes to web reference practices. The approach is to explain the need for changes, and how the suggested wPID standard can assist in achieving better practices by addressing persistency issues that are not properly addressed in current practices.

Today, there are still major issues concerning non-persistent web references. As illustration of the highly relevant need for ways to mitigate these challenges, a 2014 paper [23] found:

> … that more than 70% of the URLs within the Harvard Law Review and other journals, and 50% of the URLs within United States Supreme Court opinions, do not link to the originally cited information.

A persistent web reference is here defined as a persistent identifier (PID) for a web resource. In many cases, web references are not persistent as they consist solely of a web address and an extraction date, where the web address is a Uniform Resource Locator (URL[1]) specifying a web resource location and a mechanism for retrieving it [20]. Such references break as information on the Internet changes.

The subject of persistent web referencing has been discussed almost for as long as the web has existed. As early as 2001, a journal paper about "Persistence of Web References in Scientific Research" was published [13]. Persistent web references are needed in order to avoid the so-called "reference rot" problem, which is a combination of link rot (where a link can become inaccessible on the live web) and content decay (content changes). Examples of causes of reference rot are that a web resource has been changed moved, deleted, or placed behind a pay wall [16,18].

Persistent web referencing is increasingly relevant for research papers, as online resources are increasingly used in scholarly research (e.g. blogs) [9]. Furthermore, the persistency of web referencing is fundamental for preservation of research as well as for documentation and traceability.

There is also an increasing amount of research that is solely based on web resources [6].[2] Such research will in this paper be referred to as web research. Compared to traditional web references to documents from research papers, web researchers face a number of unique challenges, e.g. data management, references to closed archives and identification for precise annotation and referencing. However, as more and more researchers complement traditional sources with web material, these challenges will in the course of time apply to most research. Thus, when considering a general web reference proposal, the issues from web research need to be taken into account. This paper will discuss such issues within the context of Digital Humanities web research, where sustainability of web references is one of the main concerns [5].

The exact definition of a "persistent identifier" is debatable. John Kunze suggests that persistent identifier simply means that "an identifier is valid for long enough" [11]. For references in research papers this could be well over 100 years. As Juha Hakala points out: "persistent identifiers should only be assigned to resources that will be preserved for long term" [11], in other words; an identifier is worthless unless the resource it

[1] Although URL is more or less deprecated, this is the term used in the various citation templates. In order to avoid unnecessary confusion, the URL term is also used for online references

[2] Evidence can e.g. be found in reports from the BUDDAH project. See (wPID reference) wpid:archive.org:2016-03-13T011611Z:http://buddah.projects.history.ac.uk/.

points to is under a preservation program, and the identifier can be used to access the contents.

Currently, there are various approaches to the challenge of persistent web referencing, all of which includes some sort of archiving. These include registration of web resources in PID services like DOI [11], references to web archives, a method that is increasingly being applied [3], and the use of emerging citation services [16]. One of the challenges with today’s web archive reference practices is that they refer to the archive resource by including the URL for the web archives access service. This means that the archive URL may break over time due to change of access services, name shift of the archive domain or if the web archive ceases to exist [16,15].

A major obstacle to persistent web referencing for archived URLs is the temporalities not only for a single URL, but also for all the elements contained in a web page located by a URL [1]. These challenges have also been some of the motivation for the creation of the Memento protocol that can assist in finding an archived URL in a limited set of open web archives [19,2]. A recent draft report on *Interoperation Among Web Archiving Technologies* addresses these issues and points services for web archives as part of the solution [15].

The complexity of embedded material in web pages also implies that different web references can be of different quality both regarding the persistence (e.g. the trustworthiness of its survival) and its quality (a web page may not be fully harvested). Thus, in order to make trustworthy persistent references to a web page, one may need to evaluate whether several versions exist in different archives, and which version of the web page (and embedded elements) best fulfils the purpose of the reference. Therefore, this paper will include a discussion of elements to be considered when determining which web reference to use.

In order to accommodate the various challenges and support enhancement of practices for persistent web references, we propose a general global persistent web reference scheme called wPID (**w**eb **P**ersistent **ID**entifier). It is primarily focused on archived web references as a supplement to existing PID services. The wPIDs are designed to be general, global, humanly readable and agnostic regarding specific web archive implementations of access technology. The proposal is based on an analysis made from the perspectives of computer science, web collection and Digital Humanities research. Additionally, the paper describes how to represent the wPID reference scheme as a (Uniform Resource Identifier) URI scheme that can be the basis for future resolution of such identifiers [4].

The paper begins with a walkthrough of the state of the art in persistent web referencing and an introduction of the new wPID. This is followed by explanation of the various challenges in web referencing which are not covered by current best practices. Finally, the new wPID is defined as support to new best practices.

Throughout the paper the term URL will be used when addressing online web addresses (past or present), and the more general term URI will be used in relation to PID standards and archived web resources. Furthermore, any references to web resources will be provided in the new suggested wPID standard, linking to the corresponding current archive URL.

2. STATE OF THE ART AND NEW WPID

As illustrated in Figure 1, we currently have a number of different web referencing techniques and recommendations, all of which rely on the continued existence of the source material on the live web or in some sort of web archive.

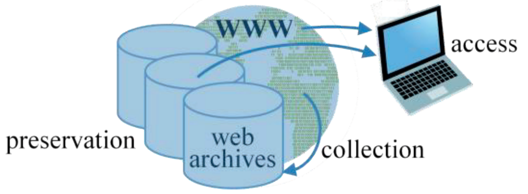


Figure 1: components for persistent references

Below, the current best known ways to make web references are described, regardless of the content that is referred. This covers the following four main ways to make references:

- ***Reference using URL and date***
  A web reference can simply consist of a web address in the live web along with a retrieval date. This is a commonly used (non-persistent) way to make web references.
- ***Reference using existing web PID services***
  A number of existing PID services provides means for content holders to register their resources. Registered web resources can then be referenced via the PID.
- ***Reference using web archives***
  Web archives can offer ways to address their content which can be used as a web reference. For example, URLs to the Internet Archive’s Wayback service.
- ***Reference using citation services***
  A number of citation services offer authors a facility to crawl and store relevant resources to be cited, where web references are provided for later access.

This section will describe these four different referencing techniques and end with a short description of the advantages and disadvantages for each technique. It will also shortly introduce the new wPID in order to compare it with current practices. Further description of the wPID will be given later in this paper.

2.1 Reference using URL and date

A commonly used web reference form is to give a URL along with its retrieval date, as for example for reference [7]:

> http://bds.sagepub.com/content/1/1/2053951714540280, retrieved March 2016

This type of reference conforms to the type of website citations using *url* and *accessdate* on “Wikipedia’s Template:Citation”[3], and is similar to most scholarly citation styles, e.g. the Chicago style[4], and the APA[5] style that both request the URL and the access date of the cited resource [16]. However, as posited, links can become inaccessible or content can change on the live web. Although commonly used, these types of references do not provide persistent identification of a resource.

2.2 Existing Web PID services

Today, there are a number of PID services offering content holders the ability to register their resources (which the content holders then preserve themselves). PID services for digital objects have been recommended as a way to ensure persistent web references [11].

[3] wpid:archive.org:2016-03-25T113243Z:https://en.wikipedia.org/wiki/Template:Citation

[4] wpid:archive.org:2015-10-07T053612Z:http://www.bibme.org/citation-guide/chicago/website

[5] wpid:archive.org:2016-03-08T233451Z:http://studygs.net/citation.htm

An example is the DOI PID service where resources can be registered and given a DOI-reference that can later be used for retrieval of the resource, e.g. the above [7] reference has the DOI reference: doi:10.1177/2053951714540280. However, PID services cannot stand alone, since many relevant references are not registered with a PID, and it is solely up to the content holder of the resources to handle registration and preservation.

A chronological list of some widespread PID services is [11]:

1. **Handle**, 1994
2. **U**niform **R**esource **N**ame (URN), 1997
3. **P**ersistent **U**niform **R**esource **L**ocators (PURL), 1995
4. **A**rchival **R**esource **K**eys (ARK), 2001

Handle[6] is a naming service that provides a mechanism for both assigning persistent identifiers to digital objects. It offers resolving of the persistent identifiers and allowing location of the authority that is in charge of the named information.

URN[7] is a concept that creates a common namespace for many different kinds of identifiers, independent of technology and location. The basic functionality of a URN is resource naming that conforms to the requirements of the general URI[8], but a URN will not impede resolution as e.g. a URL does.

PURL[9] relies on a technical solution that allows web addresses to act as permanent identifiers. It is a URL with intermediate resolution service. PURL conforms to the functional requirements of the URI, and PURL uses the HTTP protocols.

ARK[10] introduces a concept combining persistent identification and technical and administrative frameworks. This enables reference to different types of entities, e.g. agents, events and objects with metadata records. The ARK is designed to allow integration with other identifier schemes.

Besides these PID services a number of standards and services have been developed, the best known being:

1. Digital object identifier (**DOI**)
2. International Standard Book Numbering (**ISBN**)
3. National Bibliography Numbers (**NBN**)

DOI[11] makes use of the Handle System for resolving identifiers in a complete framework for managing digital objects along with policies, procedures, business models, and application tools. It is designed to be independent of the HTTP protocol.

ISBN[12] has been around as a 10 (later 13) digit Standard Book Numbering format since the 1960s. In 2001 ISBN was also described as a URN name space.

NBN[13] has no global standard, but has country-specific formats assigned by the national libraries. It is used for documents that do not have e.g. an ISBN. In 2001 NBN was described as a URN name space.

Additionally, there are communities who employ their own PID services, as for example DataCite[14] which is a community based service using DOIs for research data samples, in order to make these searchable and referable.

If a PID is registered for a resource, the idea is that the resource will be accessible through a resolver service (via live www access in Figure 1), and by a set of rules that ensures the preservation of the content that the PID addresses, but where it is the resource holder who holds responsibility for ensuring preservation program for the resource.

2.3 References to Web Archives

An increasing number of references target open web archives like Internet Archive’s collection via their Wayback service. This service offers access to a lot of the harvested web pages from the Internet Archives web archive, for example for [9]:

> https://web.archive.org/web/20160315035636/http://bds.sagepub.com/content/1/1/2053951714540280

This URL can be used as the *archiveurl* in website citations using *url*, *accessdate*, *archiveurl* and *archivedate* on the above mentioned “Wikipedia’s TemplateCitation”.

The number of Web archiving initiatives is growing. This is evident from the growth of the member list[15] of the International Internet Preservation Consortium (IIPC[16]), which is dedicated to improving the tools, standards, and best practices of web archiving for research and cultural heritage.

There is no general reference pattern for archived URLs. However, there are similar URL path patterns for web references via *archiveurl* to online web open archives using Wayback for access. All such *archiveurl* include archive date and time (denoted *date* below) and archived original URL (denoted *uri* below). The following is a list of selected open web archives and the URL patterns they use. The path differences are highlighted in bold:

- *Internet Archive (archive.org)*:
  'https://**web.**archive.org/**web**/' + <*date*> + '/' + <*uri*>
- *ArchiveIt service build by Internet Archive (archive-it.org)*
  'http://**wayback.**archive-it.org/**all**/' + <*date*> + '/' + <*uri*>
- *UK Web Archive (webarchive.org.uk):*
  'http://**www.**webarchive.org.uk/**wayback/archive**/' + <*date*> + '/' + <*uri*>
- *Portuguese web archive (arquivo.pt):*
  'http://arquivo.pt/**wayback**/' + <*date*> + '/' + <*uri*>

The differences in the paths are due to differences in the implementation of the access services at the different web archives. Thus Web references via *archiveurl* to online web archives can only be resolved as long as the web archive exists and the access path resolves to an existing access service. However, such patterns may not be valid for future access implementations.

Similar patterns may not be found for all closed archives. For example, in the Danish web archive, there are no explicit

[6] wpid:archive.org:2016-03-04T031302Z:http://handle.net/

[7] wpid:archive.org:2016-03-07T210340Z:http://tools.ietf.org/html/rfc1737

[8] URNs and URLs denote subsets of URIs [4]

[9] wpid:archive.org:2016-03-04T023751Z:https://purl.org/docs/index.html

[10] wpid:archive.org:2015-09-27T040046Z:https://confluence.ucop.edu/display/Curation/ARK

[11] wpid:archive.org:2016-03-05T022511Z:https://www.doi.org

[12] wpid:archive.org:2016-03-24T051018Z:http://www.isbn.org/ISBN_history

[13] wpid:archive.org:2016-03-31T131818Z:http://tools.ietf.org/html/rfc3188

[14] wpid:archive.org:2016-04-16T144351Z:https://www.datacite.org/about-datacite/what-do-we-do

[15] An even bigger list of web archiving initiatives can be found on wpid:archive.org:2016-03-19T171515Z:https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.

[16] wpid:archive.org:2015-04-03T190314Z:http://netpreserve.org

archiveurl, instead there is information about the placement of the resource record in a WARC file by WARC file name and offset. As the WARC file name is not part of the bit preservation and thus may change over time, researchers are recommended to supplement a reference with the archived original URL (here http://netarkivet.dk) and harvest time (given in brackets) [14]:

> http://netarkivet.dk 197800-188-20140107085943-00000-sb-prod-har-005.statsbiblioteket.dk.warc/4773261 (9:01:06 jan 7, 2014 i UTC tid)

However, this reference does not include specification of which web archive the resource was retrieved from.

Another aspect of *archiveurls* is that they may contain inherited information about special functions in web archive's access technology. An example of such a function is the *Identity* Wayback function as used by the Internet Archive. This function is called by placing 'id_' after the <date> in the *archiveurl* [17]. Another example is the function giving a snapshot image of the page[17] [16]. However, such functions may not exist in the future.

2.4 References using Citation Services

In the past years a number of citation services have emerged. These services provide on-demand archiving of a version of a given resource [16]. Examples are:

- **WebCite**[18] is an on-demand archiving system for web references. WebCite is run by a consortium and provides a tool that can archive a web reference as well as provide a new URL in the www.webcitation.org domain, where the harvested and archived referenced resource can be accessed [10].
- **archive.is**[19] (formerly archive.today) is a privately funded on-demand archiving system that takes a 'snapshot' of a web page along with a harvest of the web page (excluding active elements and scripts). The archived web page is assigned a new short URL for subsequent access.
- **perma.cc**[20] is a web reference archiving service that offers users to create a Perma.cc link where the referenced content is archived along with some metadata (URL, page title, creation date). A new link to the archived content is generated [16].

Additionally, certain web archives allow users to nominate a web page for archiving, e.g. the UK Web Archive[21], the Internet Archive[22], and Netarkivet[23]. However, for national archives like the UK Web and Netarkivet, only web pages that are considered to fall within a national scope will be archived.

Other variants exist, e.g. Zotero[24], which allow researchers to archive resources, and Wikisource that specializes in archiving Wikipidia sources[25].

2.5 References Using the New wPID

The suggested new wPID definition is a web archive reference that is independent of current web archive access technology and online access. A wPID consist of three main components, which in general are sufficient to identify any web reference in an arbitrary web archive. These three components are listed in table 1.

Table 1. Web Persistent Identifier (wPID) main parts

Part	Format	Example
Web archive	**Text**	archive.org
Date/time	**UTC timestamp**	2016-01-22T11:20:29Z
Identifier	**URI** (harvested URL)	http://www.dr.dk

For the example of reference [7], the wPID is

> wpid:archive.org:2016-03-15T035636Z:http://bds.sagepub.com/content/1/1/2053951714540280[26]

The wPID is not currently resolvable. However, it would be relatively easy to create services[27], which are based on web archive, *<date>* and *<uri>* from the wPID. This also covers closed web archives (through restricted access interface) as web archives have indexes of contents where *<date>* and *<uri>* can be use as basis for finding the current web archive URL for access.

2.6 Advantages and Disadvantages

Generally, persistency of an identifier depends on the sustainability of locating the resource by use of the identifier and that the resource content is accessible in the intended form. This is applicable to both analogue and digital resources, but the volatile nature of the Internet makes sustainability a more crucial consideration for web references. Thus, for all discussed alternatives, claims of persistency should be measured by the likelihood of a resource being locatable and accessible (with preserved contents) at a later stage.

Reference using URL and date: This reference can never be persistent. The contents can change several times during the specified date. Thus when the resource is retrieved at a later stage, there is no way to check whether it has indeed changed, and whether its contents are the intended contents.

Reference using existing web PID: Persistency relies first of all on whether a resource is registered with a PID. Of further concern is the future existence of resolver services (e.g. cases like the outage of the DOI resolver service in early 2015)[28] and whether content holders maintain the accessibility of their resource. Accessibility will rely on whether the resource holder has ensured that the resource is covered by a digital preservation program. Furthermore, for services like DOI, persistency hinges on ongoing payment of service charges. On the positive side, fees for lack of maintenance of the DOI mean that there is a strong motivation for maintaining the DOI as long as it exists.

[17] A snapshot example is: wpid:archive.org:2013-06-19T224334Z:https://archive.is/J4I1a/image.

[18] wpid:archive.org:2016-03-06T000304Z:http://webcitation.org/

[19] wpid:archive.org:2016-02-19T153542Z:http://archive.is/

[20] wpid:archive.org:2016-03-05T093301Z:https://perma.cc/

[21] wpid:archive.org:2016-03-04T052011Z:http://www.webarchive.org.uk/ukwa/info/nominate

[22] wpid:archive.org:2016-03-01T085607Z:http://archive.org/web/

[23] wpid:archive.org:2016-03-03T220018Z:http://netarkivet.dk/

[24] wpid:archive.org:2016-03-06T080434Z:https://www.zotero.org/

[25] wpid:archive.org:2016-02-27T212014Z:https://en.wikipedia.org/wiki/Wikisource

[26] Omission of ":" in date/time is described later.

[27] Discussion on APIs (including Open Wayback) includes mentioning of APIs for such services [15].

[28] wpid:archive.org:2016-03-10T044938Z/http://blog.crossref.org/2015/03/january-2015-doi-outage-followup-report.html.

Reference using Web Archives: Persistency relies on the continued existence of a web archive, and the preservation program that the archive has for its resources. As mentioned in [16]: "one link rot problem was replaced by another" if the archive ceases to exist. Furthermore, future existence of compatible access services as archive links with inherited service and service parameters may be at risk due to future changes in access tools or archive ownership.

Reference using Citation services: These services are in many aspects similar to web archives, and so the persistency of references depends on the continuation of the given service and the future existence of compatible access services as well as preservation program for the resources. An example of a vanished citation service is the former mummify.it citation service mentioned in [16], which in the Internet Archives web archive was used in the period from 2013-08-30 to 2014-02-14. In 2015, it had changed to an online shoe shop and is now inactive.

Reference using the new wPID: As for web archive and citation services references persistency rely on the existence of the web archive and its preservation program. The advantage is that a wPID has sufficient information to identify a referred source in any web archive independent on access implementations and/or generated IDs like shortened URLs. Current lack of resolving may be seen as a disadvantage, but services can easily be made and these services can be maintained to point to access platforms as they change due to change in technologies.

Logical preservation of resources needs to be part of the required preservation program for all resources pointed to by persistent web references. Logical preservation covers aspects of keeping the resource accessible in spite of technology and format changes. For controlled web resources (e.g. handled by PID systems) this can include migration of formats. One of the solutions for web archives that is now being investigated is emulation, e.g. oldweb.today.[29]

It should be noted that a major difference between PID services and web archives is the placement of responsibility of preservation management. PID services only provide identifiers where the resource holders are responsible for content preservation, while it is the web archives that have this responsibility for archive references (which is the same for most citation services).

In the rest of this paper, we will leave out further analysis of the ***URL and date*** reference type, as it can never become a persistent way of referencing a web resource. As the aim here is to focus on references to the archived web as a supplement to existing practices, where there may not be a holder of the resource, further analysis of ***existing Web PID services*** is also left out.

The focus in the rest of the paper will be on what a web reference actually means, taken into account the needs from researchers, the quality of a web reference according to its purpose and the ambiguities that can be inherited in a web reference.

3. RESEARCH IN WEB MATERIAL

In many ways, web researchers using web references face challenges that are similar to referencing to digital papers and resources. However, in the field of web research it is more obvious that there are additional requirements, which must be taken into account in order to make the best possible proposal for general web references.

Here, the additional web research requirements are illustrated by investigating current issues in Digital Humanities. Today, Digital Humanities is used to describe at least two entwined processes:

1. With the advent of new computational techniques researchers are able to process a hitherto unseen amount of data (whether born-digital or reformatted), and
2. As the hegemony of conventional media is being challenged, scholars must now trace a number of cultural processes and interactions in the form of digital artefacts [8]

These new circumstances call for new measures, yet the lack of a shared and stringent methodology is a well-recognized cause for concern within Digital Humanities [6,7]. This is particularly true when it comes to research in web materials – a budding empirical field within both the Social Sciences and the Humanities. Web researchers have to cope with unique issues due to the dynamic content of their empirical field.

Web research, whether using the live or archived web, is faced with challenges related to both data management and identification for annotation and referencing (figure 2). Identification is here understood both as the actual search for material as well as the means to identify the precise content of a web reference.

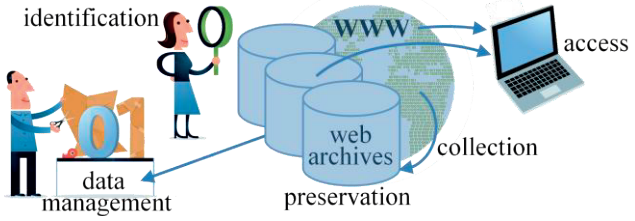


Figure 2: components for web research

It should be noted that a lot of such references have a potential problem regarding access, as usage of non-public references might be restricted (e.g. denied or limited to on-site users) due to regional legal frameworks.

3.1 Analogue Standards and Current Needs

From the perspective of an institution dedicated to cultural heritage preservation and research, this paper grapples with one of the cornerstones of sound methodology, namely the ability to give citations in keeping with current scientific standards.

The purpose of accurate referencing is – first and foremost – to give readers an opportunity to assess the reliability and provenance of a given source material as well as to retrace and reproduce research steps. As such, the current inability to provide reliable references touches on all components – identification, data management and access – for web research (figure 2).

Within the Humanities and Social Sciences, reference systems are structured to provide the most accurate link to a given object. In original sources this entails pointing to distinct passages, foot notes, a word or even marginalia. For published material, which can appear in a number of different editions, citation styles often require users to include unique identifiers

[29] wpid:archive.org:2016-03-08T205232Z:http://oldweb.today/

such as the former mentioned ISBN and the related Serial Item and Contribution Identifier (SICI[30]) for periodicals.

Yet for web pages, the most commonly used style guides (e.g. formerly mentioned Chicago style) request nothing beyond the URL and a date indicating the date a URL was "last modified" or merely "accessed".

The discrepancy between these standards and the requirements of conventional research means that researchers might shy away from incorporating web materials or that web research will in itself be discredited due to methodological inadequacies.

In conclusion, there is a present and urgent need for a persistent web referencing scheme on par with that for analogue materials.

4. WEB REFERENCING CHALLENGES

The differences between referencing scheme for analogue and web references are mainly due to the dynamic nature of the web and the temporalities within complex web pages. The differences and related challenges are especially pertinent for researchers referring to complex web resources, as is often done in web research.

Determining whether a link is "alive" or "dead" poses an additional challenge. There are notions of dead links in connection with a citation that points to the live web, but there is no clear definition of what "dead" entails if we take into account that a link can potentially live on in an – possibly off-line – archive. Since persistency does not necessarily rely on what is online, this needs to be taken into account in regards to the challenges of persistent web references.

The following sections describe the dynamics and context of "dead" links, and are concluded by a section discussing the quality of a persistent web reference with respect to these issues.

4.1 A Relative Whole with Temporalities

One of the major challenges with persistent references of web pages is that these are mostly comprised of separate parts as illustrated in Figure 3. In this example the URL only refers to an html element, which includes style sheets, text, links etc. The links are new URLs to elements embedded in the web page (e.g. images) or URLs to elements in form of other resources (e.g. link to PDF files).

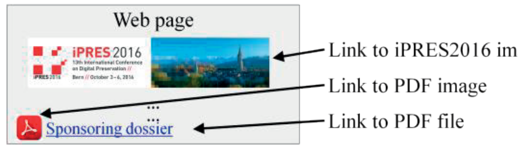


Figure 3: Elements in a web page

Different elements are harvested at different times, and some elements may only partially be harvested or not at all. This causes troublesome temporalities or incompleteness in the web archives.

For single self-embedded elements like a PDF file the temporalities are not an issue. However, for web pages with dynamic contents the temporalities can be crucial, see for example slide 8 of the *Evaluating the Temporal Coherence of archived pages* IIPC presentation [1] in which a weather forecast predicting storm and rain is depicted with a clear radar image extracted 9 months after the harvest of the main page.

The challenges of temporalities and coverage make web archives a rather difficult academic resource [6].

The temporality challenge implies that an archived web reference may be ambiguous. Traditionally, it is the archive software that picks the version of page elements, but for an exact research reference it may be necessary to specify each of the elements. Consequently, all parts should be denoted with wPIDs, which in some cases may incorporate wPIDs for parts found in separate web archives (also mentioned in [15]).

Another challenge is that web archives – open as well as closed – will never be able to contain snapshots of the entire Internet. One reason is the continuous change in content and the challenge of temporality, but also the fact that the amount of data is simply too big. Today, a number of web archives cover different parts of the web. Typically, national web archives systematically harvest the Top Level Domains of the country, but Top Level Domains like .com, .org and .nu are not covered in full by any web archives.

4.2 Variety of Errors in Web Page Search

When looking for or looking up a web reference in a web archive, it is important to be aware of the possible reasons why a page seems to be missing from the archive.

In general, a "missing" reference can either be caused by limitations or errors in how the related URL was collected or how it is accessed. This is illustrated in Figure 4.

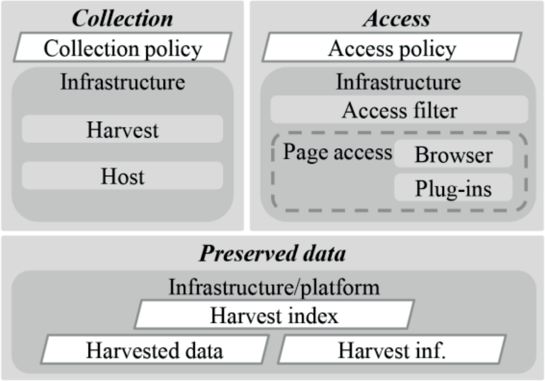


Figure 4 Web Archive Infrastructure

Below, each of the three components is described with its name in ***bold/italic***, and subcomponents are highlighted by being written in *italic*. Archived URLs are denoted URIs, as they exist within a web archive and therefore do not represent locators. The description includes all possible error sources, including sources that do not exist in the web archive.

Collection causing a missing web reference may be due to:

Collection Policy: In case the web reference is not covered by policy and thus never collected by the web archive in question.

Harvest of a host web reference can fail for a number of reasons:

- Errors in infrastructure (e.g. missing network connection)
- Bad harvest settings (e.g. stopped by max bytes)
- Cannot harvest inherited material (e.g. Flash)
- Cannot harvest scripting (e.g. java scripts and AJAX)
- Harvester fails (e.g. due to crawler traps) or was killed
- Host or part of host is down or unavailable
- Host does not allow the harvest

In most cases, the above harvesting errors mean that a reference is not usable.

Access causing a missing web reference may be due to:

Access policy enforced by an *Access filter*: In some case there is limited access, e.g. respecting robot.txt by disallowing access or filtering of illegal material, special granted access to the web archive may be required to check if the reference is correct.

Page access can fail for a number of reasons:

- Unavailable preserved data due to infrastructure problems (e.g. network or access application is down)
- Errors or limitations in access program (e.g. cannot show https & ftp URIs or URIs with special characters like æ, ñ)
- Misunderstood date/time as it is specified as UTC time-stamp or errors in local conversion of time
- Errors in the index used for look up of data (e.g. wrongly generated index or problems with de-duplication)
- Normalization of URI doesn't conform with indexed URI
- Access programs may interfere with the display[31]

Browser used for access does not render the page correctly

- because the browser does not comply with standards used (or exceptions from standards)
- because the web page is from a time period requiring special browsers (e.g. Netscape) or special versions of browser

Plug-ins needed for access do not exist or are not supported on rendering platform (or no longer supported).

Furthermore, the ***Preservation data*** can cause access errors, either by having an erroneous *Harvest index*, by errors in *Harvested data* (e.g. packed with wrong ARC file offset) or by *Infrastructure / platform errors* (e.g. server with preserved data is down).

Understanding these potential error sources, it is now possible to classify whether a link has truly died, and what sort of "death" we are encountering.

4.3 Different types of "Dead" links

A 'dead link' is commonly associated with link or reference rot, however, there are many ways that a link can 'die', therefore we need to look closer at the variations of what link rot means.

The archived content for a URI depends on harvest and consequently resolving of the URLs. Thus, a proper analysis of a persistent web reference must include consideration of the different types of "deaths" of both web URLs and archived URIs.

The following description relates to the search for a web reference in the form of a URL/URI (and possible date/time) with *expected* contents and may refer to HTTP codes[32] resulting from URL/URI requests.[33]

The following lists possible types of "deaths" for URLs on the live web:

- *Indistinguishably dead*: The page does not seem to exist, e.g. HTTP return a code indicating net or host errors
- *Instantly dead*: The page cannot be found, e.g. it resolves to an HTTP 404 code "page not found" generated by the server or a web page offering you to buy the domain or just redirects to some random domain.
- *Identity dead*: the page is not the expected page, e.g. due to new domain owner.
- *Simulated dead*: the page cannot be accessed due to some sort of blocking such as content filters or firewalls (also called soft errors in [15]).

Note that the classifications are conceptual and cannot necessarily be linked to specific technical traceable HTTP codes. This means that it can be hard to verify whether a page is *Instantly* or *Indistinguishably dead*. For example, a disappeared domain can resolve with the same error as missing network connection.

Today, live link checkers can search for dead links mainly by relying on technical HTTP codes. That means a "page not found" generated internally from a server may be regarded as a successful unbroken link as it will not return an HTTP 404 code. *Identity dead* links will also be reported alive and the link checker will not be able to determine whether a link is *Indistinguishable dead* or *Instantly dead*.

It becomes even more complicated when searching for content in an archive due to the possible harvest/access/preservation errors described in the previous section:

- *Archival dead*: A URI (and date/time) doesn't exist in the archive
- *Partially dead*: A URI (and date/time) does exist in the archive, but cannot be correctly displayed because pieces are missing due to harvest limitations
- *Technology dead*: A URI (and date/time) does exist in the archive, but it is not correctly displayed because of access limitations, e.g. due to browser or plug-in limitations
- *Apparently dead*: A URI (and date/time) cannot be found in the archive, due to errors in the access part, e.g. cannot access HTTPS URIs, wrong indexes etc.
- *Temporarily dead*: A URI (and date/time) can be found in the archive, but infrastructure problems or limitations like robot.txt make it temporarily inaccessible

Again these death types are conceptual classifications, and they are not necessarily easy to recognize, as symptoms of errors may differ for different access applications.

Finally, there is the *Ultimate dead* meaning that the URL/URI is neither in any archives nor on the live web. This will probably be impossible to verify, as we can never be sure whether we know all archives and whether all possible errors are taken into account.

4.4 Quality of a Persistent Web Reference

In general, use of web references as part of research or articles needs to be carefully evaluated for the intended purpose of the reference and its persistency quality both for the identifier and the resources identified. Specifically, for ***Reference using Web Archives*** the various mentioned web referencing challenges should be taken into account.

When choosing a web reference, the first task is to ***identify*** the needed reference in a web archive (or citation service) and verify that the resource can be accessed. For example, the reference is not *Apparently dead*, e.g. due to errors in the access application, and it is not *Instantly dead*, because of

[30] wpid:archive.org:2016-03-04T102536Z:http://www.niso.org/apps/group_public/download.php/6514/Serial%20Item%20and%20Contribution%20Identifier%20(SICI).pdf

[31] In the case of the Internet Archive Wayback, however, there are options to mitigate this challenge [17].

[32] wpid:archive.org:2016-02-29T024353Z:https://en.wikipedia.org/wiki/List_of_HTTP_status_codes

[33] As noted in [18] contents from a dead link can be provided by other channels, e.g. contacting the author of a thesis referred.

reconstruction of the web archives access platform and the fact that the resource therefore needs to be found under another URL.

The next task would be to evaluate the referred ***contents*** with respect to referencing purpose. For example, it is not a case of *Simulated dead*, e.g. that the harvested resource is not just a login screen for password protected content.

Furthermore, it must be checked that the referred resource is of the right *Identity*, as could be the case for the mentioned mummify.it example, which at one stage was a citation service and at another stage a shoe sales site. In this example it is easy to recognize, but differences may be subtle and thus harder to recognize.

The purpose of the reference is crucial, since *Partially dead* referred content may fulfil its purpose, e.g. a web page containing complicated java script and flash can be harvested incompletely, yet the rest of the content might still be accessible and adequate for the referencing purpose [18].

Finally, an evaluation of the ***persistency*** should be performed in terms of future accessibility of the resource. This includes evaluation of the identifier as well as the contents referred.

The referenced resource may suffer *Archival dead* if the web archive partly or fully ceases to exist, i.e. an evaluation of the sustainability of the web archive(s) should be included. As an example, this paper will have a lot of invalid wPIDs in the future if the Internet Archive web archive is shut down.

The referenced resource can suffer a *Technology dead* if the web archive does not have a proper preservation program, and thus fails to keep the resource's existence or resource's functionalities available over time. Sustainability of access services should also be evaluated, in particular for web archives in the form of citation services relying on shortened URLs as persistent identifiers. Business and funding models are crucial elements in this evaluation.

5. SUGGESTED WPID REFERENCES

The suggested wPID aims at simplicity, readability, sustainability and transparency. The definition is based on analysis of the state of the art of persistent referencing; relevant web standards and the need for web research and the various challenges described in the previous sections. Furthermore, it takes into account that it could benefit from becoming an accepted permanent URI scheme [4] as described and explained in the last part of this section.

5.1 General wPID Definition Suggestion

As described in Table 1, the wPID consists of three main parts. Below, there are provided more details on choices made for their structure and how this relates to existing web standards like the WARC standard (packaging format used for many web archives) [12] and URI scheme standard [4].

- **Web archive**
  Is specified by Sequence of URI Unreserved Characters ('-', '_', '.', '~', alpha: 'a'-'z', 'A'-'Z' or digits: '0'-'9'.

- **Date/time**
  Is specified as a short UTC timestamp with the same definition as the WARC-Date field in the WARC standard, i.e. formatted as YYYY-MM-DDThh:mm:ssZ, conforming to the W3C profile of ISO 8601 [12,22], but omitting ":" in order to conform with the URI standard (as explained later).

- **Identifier**
  Is a URI as defined for the WARC-Target-URI field in the WARC standard. This field is for the archived URI which must have a value written as specified in [4,12]

There are no real restrictions to what a web archive name can be. In the examples used in this paper, the domain name for the archive is used. The reason for this is that the domain names are known today. However, proper names could be used if a register is created (similar to the NAAN registry[34] for ARK) and possibly maintained by the IIPC or a similar body. Such names could be *InternetArchive* for archive.org, DKWebArchive for the Danish web archive etc. In all cases, a register should be made at some stage, since archive domains can change (e.g. archive.today is now named archive.is). Note that such a registry should allow several names for each archive, since archives may be merged or renamed. Thus, old references need to remain persistent and traceable, regardless of use of the old name.

Additionally, we need to be able to avoid the ambiguity of *the parts and the whole*. We can accomplish that by specifying a *contentspec* parameter, which can have the values:

- *harvest*, in case the parts are taken from the archive in the traditional way,

- *part*, in case the wPID is to be interpreted as the single web page part.

We assume that "harvest" is default in case nothing else is specified.

Finally, in order to make it compatible with a URI, it must follow the URI syntax [4] and be defined as a URI scheme[35]. The URI syntax causes some challenges, since the definition will be recursive, as the defined wPID URI contains a URI in its definition:[36]

*wpid-**URI*** *= scheme ":"*
*<hierarchical part incl. archived-**URI**>*
["?" query] ["#" fragment]

The challenge is that there is no way to distinguish whether queries and fragments belong to the *wpid-**URI*** or the *archived-**URI.*** Thus queries and fragments cannot be given unambiguously to the *wpid-**URI***. The information about the *contentspec* therefore cannot be specified as a query, but instead it needs to be part of the *hierarchical part*. There is already an indirectly proposed solution for dealing with this challenge. Internet Archive specifies the access parameters for the Wayback, as previously explained, by adding a flag to the timestamp portion. Thus, the challenge can be solved by having the suggested *contentspec* as timestamp flag extensions in the same way.

Another challenge with the URI syntax is the limitation on the use of delimiters within the *hierarchical part*. If we define the wPID as a URI with *scheme* "wpid" and a *hierarchical part* as a *path* with no authority and without segments, then the best choice of delimiter is ":". However, this collides with the colons used in the UTC timestamp. The suggestion to work around this challenge is to strip the colons in the UTC timestamp.

[34] wpid:archive.org:2015-09-17T131414Z:http://www.cdlib.org/uc3/naan_table.html

[35] The wPID URI scheme is registered as a provisional URI scheme, see wpid:archive.org:2016-04-17T062512Z:http://www.iana.org/assignments/uri-schemes/uri-schemes.xhtml and wpid:archive.org:2016-04-17T062702:http://www.iana.org/assignments/uri-schemes/prov/wpi.

[36] The syntax is defined using the same symbols as other RFC standards, i.e. test in double quotes "" are text to be written, text without double quotes are entities defined later, '/' means 'or', '[' and ']' surrounds optional parts, '(' and ')' surrounds parts that can be repeated where '+' means one or more times, and finally '<' and '>' surrounds explanatory text.

The resulting wPID definition is consequently the following:

```
wpid        = "wpid:" webarchive ":" archivedate
              [ contentspec ] ":" archiveduri

webarchive  = +( unreserved )
contentspec = "harvest_" / "part_"
archivedate = <as date/time in table 1 stripped for ":">
archiveduri = <as identifier in table 1>
unreserved  = <as defined in RFC 3986 [4]>
```

A wPID (for an archive context) consisting of the example elements from table 1 would then be:

wpid:archive.org:2016-01-22T112029Zharvest_:
http://www.dr.dk

since *harvest* is the default *contentspec* this is the same as

wpid:archive.org:2016-01-22T112029Z:http://www.dr.dk

Note that a wPID cannot leave out any of the syntax components from table 1, since all will be needed in order to make a persistent identifier. Thus the wPID should only be used when the reference is verified to be present in the specified archive.

The analysis of the quality of traditional web references suggests a need to add additional information about the reference target quality. However, it is not possible to do an analysis that can cover all possible scenarios, and it doesn't add any additional value on how to find the resource, thus this is not a subject for standardization, but could instead be made as a comment along with a wPID reference.

5.2 Why define wPIDs as URIs

It may not seem obvious why the wPID has to be defined as a permanent URI scheme in the form of a Request for Comments (RFC) as part of publication from the Internet Engineering Task Force (IETF)[37]. The claim here is that the benefits are worthwhile in spite of the disadvantages in form of the (not very elegant) workarounds for parameter and delimiters.

The benefits of a new wPID URI schema are first of all that it is a standard for the World Wide Web deployed since the creation of the Web [21]; secondly, it is the next step towards possible creation of some sort of resolving service via a browser, accessing locally or globally. For example, tools like the Memento tool could assist in wPID resolution, or special browser plug-ins recognizing wPID URIs could redirect to current access implementation (or APIs) using the HTTP/HTTPS protocols, and likewise from local browsers to closed archives.

6. DISCUSSION & FUTURE WORK

True persistence of a web reference will always come down to the existence of the archive responsible for the preservation of the reference contents and accessibility. This is true for all archive material, but will probably be a bigger issue for web archives as their existence hinges on the legislation and/or business models, which they are grounded on.

There are still challenges that are not fully addressed concerning data management, including corpus building and annotation. Some of the challenges relate to having unambiguous references web pages that may consist of several web elements that originate from one or more web archives. These challenges will be the basis for further investigation within current research projects[38] based on the suggested wPID standard.

Search for the right web reference has not been the focus of this paper. However, it is needed, and the Memento protocol is well suited for this task at least for the open web archives covered.

Additionally, when choosing a web reference in a web archive, it is important to take into account the possible temporalities and the evaluation of persistency of the archives. It will be the user of the web reference that is responsible for such evaluations, but compilation of guidelines for this task could be useful.

It will also be worth considering whether wPIDs could be applied as persistent references to digital library resources in general, i.e.

wpid:<library domain>:<timestamp>:<UUID for resource>

could be a reference to a library resource registered with a *UUID* and archived at the time specified in the *timestamp*. In this way it would also be possible to distinguish persistent identifiers for original versions and migrated versions of resources.

7. CONCLUSION

We have argued that there is an urgent need for better persistent web referencing practices, in order for researchers to include valid and precise web references in their research.

We proposed a new best practice for web referencing with a supplementary new wPID standard for references to web archives.

The paper has included a selected number of challenges within today's practices and future references and we have made a walkthrough of issues to be aware of when choosing a persistent web reference scheme. In particular, for wPIDs, this includes thorough validation of the web reference by the users of the reference before using it, as well as sustainability of the web archive, its preservation program for web resources and ability to offer access services based on archived URI and harvest time.

In addition, we have argued for the benefits of defining the wPID as an RFC standard by defining it as a URI scheme. This opens up the opportunity for a standard that can be used for technology independent access to web archives in the future.

The paper has included illustrations of the complexity and ambiguity that has become part of today's web referencing practices, especially for references to complex web pages. We argued that the suggested standard can be the basis for further studies of how to cope with these challenges including data management in web research.

8. ACKNOWLEDGMENTS

The authors wish to thank colleagues from the Royal Library of Denmark, especially Jakob Moesgaard, who contributed with web collection expertise. Also thanks to Jane Winters from University of London for her useful comments. Finally, thanks to digitalbevaring.dk for the illustrations.

[37] wpid:archive.org:2016-03-27T010831Z:https://www.ietf.org/

[38] The two co-authors of this paper are each conducting their Digital Humanities web research projects partly financed by the Danish Ministry of Culture: One project documenting online mourning and memorial culture and one project about the archiving of the various and constantly changing digital platforms of the literary field.

9. REFERENCES

[1] Ainsworth, S. G., Nelson, M. L., Van De Sompel, H. 2015. *Evaluating the Temporal Coherence of archived pages*, wpid:webarchive.org.uk:2016-01-14T233144Z:http://netpreserve.org/sites/default/files/attachments/2015_IIPC-GA_Slides_18_Nelson.pptx.

[2] Alam, S., Nelson, M.L., Van de Sompel, H., Balakireva, L.L., Shankar, H., Rosenthal, D.S.H. 2015. *Web Archive Profiling Through CDX Summarization*, wpid:archive.org:2015-12-05T065734Z:http://www.cs.odu.edu/~mln/pubs/tpdl-2015/tpdl-2015-profiling.pdf#sthash.G2BjVUf4.dpuf.

[3] AlNoamany, Y., AlSum, A., Weigle, M.C., Nelson, M.L. 2014. *Who and what links to the Internet Archive*. In Research and Advanced Technology for Digital Libraries, Volume 8092, Lecture Notes in Computer Science, Springer pp. 346-357.

[4] Berners-Lee, T., Fielding, R., Masinter, L. 2005. *Uniform Resource Identifier (URI): Generic Syntax* (RFC 3986), wpid:archive.org:2016-03-26T121040Z:http://www.ietf.org/rfc/rfc3986.txt.

[5] Blaney, J. 2013. The Problem of Citation in the Digital Humanities, wpid:archive.org:2015-04-29T220653Z:http://w ww.hrionline.ac.uk/openbook/chapter/dhc2012-blaney.

[6] Brügger, N., Finnemann, N.O. 2013. *The Web and Digital Humanities: Theoretical and Methodological Concerns*. Journal of Broadcasting & Electronic Media 57, nr. 1, pp. 66–80. doi:10.1080/08838151.2012.761699.

[7] Burrows, R., M. Savage. 2014. *After the Crisis? Big Data and the Methodological Challenges of Empirical Sociology*. Big Data & Society 1, no. 1, doi:10.1177/2053951714540280.

[8] Clavert, F., Serge, N. 2013. *Digital Humanities and History. a New Field for Historians in the Digital Age*. In L'histoire contemporaine à l'ère numérique / Contemporary History in the Digital Age, Clavert F., Serge N (eds.), pp. 15–26.

[9] Davis, R. M. 2010, *Moving Targets: Web Preservation and Reference Management*. Ariadne Web Magazine, issue 62, wpid:archive.org:2016-03-22T034748Z:http://www.ariadne.ac.uk/issue62/davis/.

[10] Eysenbach, G. 2005. *Going, Going, Still There: Using the WebCite Service to Permanently Archive Cited Web Pages*. JMIR publications, Vol 7, No 5.

[11] Hakala, J. 2010, *Persistent identifiers – an overview*, wpid:archive-it.org: 2013-10-10T181152Z:http://metadaten-twr.org/2010/10/13/persistent-identifiers-an-overview/.

[12] ISO 28500:2009. 2009, *WARC (Web ARChive) file format*.

[13] Lawrence, S., Pennock, D. M., Flake, G. W., Krovetz, R., Coetzee, F. M., Glover, E., Nielsen, F. Å., Kruger, A., Giles, C. L. 2001. *Persistence of Web References in Scientific Research*, wpid:archive.org:2016-03-05T222400Z:http://clgiles.ist.psu.edu/papers/Computer-2001-web-references.pdf.

[14] Netarkivet. Brugermanual til Netarkivet (User Manual for the Netarkivet). 2015. wpid:archive.org:2016-03-10T143320Z:http://netarkivet.dk/wp-content/uploads/2015/10/Brugervejledning_v_2_okt2015.pdf.

[15] Rosenthal, D.S.H., Taylor, N., Bailey, J.: *DRAFT: Interoperation Among Web Archiving Technologies*, wpid:archive-it.org: 2016-04-13T114517Z:http://www.lockss.org/tmp/Interoperation2016.pdf.

[16] Van de Sompel, H., Klein, M., Sanderson R, Nelson, ML. 2014. *Thoughts on Referencing, Linking, Reference Rot*, wpid:archive.org:2016-03-03T190515Z:http://mementoweb.org/missing-link/.

[17] Wikipedia. 2016. *Help: Using the Wayback Machine*, wpid:archive.org:2016-03-12T041354Z:https://en.wikipedia.org/wiki/Help:Using_the_Wayback_Machine.

[18] Wikipedia. 2016. *Link rot* wpid:archive.org:2016-03-19T113853Z:https://en.wikipedia.org/wiki/Wikipedia:Link_rot.

[19] The Memento Project, *About the Time Travel Service*, wpid:archive.org:2016-03-15T080039Z:http://timetravel.mementoweb.org/about/.

[20] Wikipedia. 2016. *Uniform Resource Locator*, wpid:archive.org:2016-04-05T134446Z:https://en.wikipedia.org/wiki/Uniform_Resource_Locator.

[21] World Wide Web Consortium (W3C). 2004. *Architecture of the World Wide Web, Volume One*. wpid:archive.org:2016-04-04T213830Z:https://www.w3.org/TR/webarch.

[22] World Wide Web Consortium (W3C). 1997. *Date and Time Formats*. (W3C profile of ISO 8601). wpid:archive-it.org:2016-03-31T232655Z:http://www.w3.org/TR/NOTE-datetime.

[23] Zittrain, J., Albert, K., Lessig, L. 2014. *Perma: Scoping and Addressing the Problem of Link and Reference Rot in Legal Citations*, doi:10.1017/S1472669614000255.

Preserving Websites Of Research & Development Projects

Daniel Bicho
Foundation for Science and Technology:
Arquivo.pt
Av. do Brasil, 101
1700-066 Lisboa, Portugal
daniel.bicho@fccn.pt

Daniel Gomes
Foundation for Science and Technology:
Arquivo.pt
Av. do Brasil, 101
1700-066 Lisboa, Portugal
daniel.gomes@fccn.pt

ABSTRACT

Research and Development (R&D) websites often provide valuable and unique information such as software used in experiments, test data sets, gray literature, news or dissemination materials. However, these sites frequently become inactive after the project ends. For instance, only 7% of the project URLs for the FP4 work programme (1994-1998) were still active in 2015. This study describes a pragmatic methodology that enables the automatic identification and preservation of R&D project websites. It combines open data sets with free search services so that it can be immediately applied even in contexts with very limited resources available. The "CORDIS EU research projects under FP7 dataset" provides information about R&D projects funded by the European Union during the FP7 work programme. It is publicly available at the European Union Open Data Portal. However, this dataset is incomplete regarding the project URL information. We applied our proposed methodology to the FP7 dataset and improved the completeness of the FP7 dataset by 86.6% regarding the project URLs information. Using these 20 429 new project URLs as starting point, we collected and preserved 10 449 947 Web files, fulfilling a total of 1.4 TB of information related to R&D activities. All the outputs from this study are publicly available [16], including the CORDIS dataset updated with our newly found project URLs.

Keywords

Automatic identification; Preservation; Web archives; Research and Development projects

1. INTRODUCTION

Most current Research & Development (R&D) projects rely on their websites to publish valuable information about their activities and achievements. However, these sites quickly vanish after the project funding ends. During the funding work programme FP7 the European Union invested a total of 59 107 million EUROS on R&D projects. Scientific outputs from this significant investment were disseminated online through R&D project websites. Moreover, part of the funding was invested in the development of the project websites themselves. However, these websites and the information they provide typically disappear a few years after the end of the projects. Websites of R&D projects must be preserved because:

- They publish valuable scientific outputs;
- They are aggregators of scientific outputs related to a given theme because the R&D projects are typically funded in response to a call on proposals to solve specific societal or scientific problems;
- They are not being exhaustively preserved by any institution;
- They are highly transient, typically vanishing shortly after the project funding ends;
- They constitute a trans-national, multi-lingual and cross-field set of historical web data for researchers (e.g. social scientists).

The constant deactivation of websites that publish and disseminate the scientific outputs originated from R&D projects causes a permanent loss of valuable information to Human knowledge from a societal and scientific perspective. Web archiving provides a solution to this problem. Web archives can preserve this valuable information. Moreover, funding management datasets can be enriched with references of the preserved versions of the project websites that disappeared from the live-Web. However, websites that publish information related to R&D projects must be firstly identified so that web archives can preserve them.

There has been a growing effort of the European Union, and governments in general, to improve transparency by providing open data about their activities and outputs of the granted fundings. The European Union Open Data Portal [8] is an example of this effort. It conveys information about European Union funded projects such as the project name, start and end dates, subject, budget or project URL. Almost all this information is persistent and usable through time after the project or funding instruments end. The exception is the project URL. As websites typically disappear a few years after their creation [31], the R&D management databases available at The European Union Open Data Portal, such as the datasets of the CORDIS EU research projects, suffer degradation by referencing complementary online resources that became unavailable and were not systematically preserved neither by the funder nor the funded entities. Moreover, the CORDIS EU research project datasets have incomplete information regarding the projects URLs. From a total of 25 608 project entries, only 2 092 had the project URL field filled. Thus, about 92% of project websites could not be identified and therefore their preservation was challenged.

The Foundation for Science and Technology (FCT) [10] is the official Portuguese institution that manages research funding and e-infrastructures. Arquivo.pt [2] - the Portuguese Web Archive is one of the research infrastructures managed by FCT and its main objective is to preserve web material to support research activities. Hence, the websites of R&D projects are priority targets to be preserved. The objective of our work was to study techniques to automatically identify and preserve R&D project websites funded by the European Union based on existing free tools and public data sets so that they can be directly applied by most organizations and information science professionals, without requiring the intervention of computer scientists, or demanding computing resources (e.g. servers, bandwidth, disk space). The main contributions of this work are:

- Quantitative measurements about the ephemera of EU-funded project websites and their preservation by web archives;
- A test collection and methodology to evaluate heuristics to automatically identify R&D project websites;
- A comparative analysis between heuristics to automatically identify URLs of R&D projects using free search services and publicly available information datasets;
- A list of web addresses of existing R&D project sites that can be used by web archives to preserve these sites or by management institutions to complement their datasets.

We believe that the results described here can be immediately applied to bootstrap the preservation of EU-funded project websites and minimize the loss of the valuable information they convey as has been occurring for the past 22 years.

2. RELATED WORK

The vastness of the web represents a big challenge with regard to preservation activities. Since it's practically impossible to preserve every web content, the question remains: "how much of the web is archived? [20]". The problem of link rot is a serious and prevalent problem that jeopardizes the credibility and quality of scientific literature that increasingly references complementary online resources essential to enable the reproducibility of the published scientific results (e.g. experimental data). A study about the decay and half-life period of online citations cited in open access journals showed that 24.58% of articles had online citations and 30.56% of them were not accessible [41]. The half-life of online citations was computed to be approximately 11.5 and 9.07 years in Science and Social science journal articles respectively. However, the link rot problem in scientific publications is not a problem of open access journals. The unavailability of online supplementary scientific information was also observed across articles published in major journals [28, 30]. The problem of link rot is cross-field and has been scientifically reported over time. For instance, it was observed among scientific publications in the fields of Computer Science in 2003 [44], Information Science [45] in 2011 or Agriculture in 2013 [43]. We believe that many of the link rot citations reference resources published on project websites that meanwhile became unavailable. Preserving these sites would significantly contribute to maintain the quality of scientific literature.

Since the early days of the web, several studies addressed the problem of identifying relevant web resources. Focused crawling approaches try to identify valuable information about a specific topic [25]. ARCOMEM - From collect-all archives to community memories was a EU-funded research project conducted between 2011 and 2013 that aimed to study automatic techniques to identify and preserve relevant information regarding given topics specially from social media. Ironically, the project website is no longer available and could only be found in publicly available web archive [22]. ARCOMEM studied, for instance, how to perform intelligent and adaptive crawling of web applications for web archiving [29] or how to exploit the social and semantic web for guided web archiving [39]. However, implementation of such approaches is too complex and entails a significant amount of resources, requering powerful crawlers and bandwidth resources to harvest the web looking for relevant resources. The process can be optimized but considering the dimensions of web data, it is still too demanding to be implementable by most Cultural Heritage Organizations. web services, such as live-web search engines, have already crawled and processed large amounts of web data, and provide search services to explore it. Bing Web Search API [3] and Google Custom Search API [11] are examples of commercial APIs that can be used to explore those web data. However, these services limit the number of queries per user based on the subscribed plan. Contrarily, non-commercial APIs like Faroo [9] don't have limitations on the number of queries a user can perform, but the search results tend to be worse due to the relatively low amount of web data indexed.

Therefore, alternative approaches that explore existing services and resources to identify and preserve relevant web content have been researched. Martin Klein and Michael Nelson proposed methods to rediscover missing web pages automatically through the *web Infrastructure* [33]. In their study they have *a priori* information about the original URL which they used it to build several heuristics to rediscover the missing web pages. Shipman et al. used page titles to rediscover lost web pages referenced on the DMOZ web directory by using the Yahoo search engine [42].

Websites containing information regarding European Union fundings and R&D projects are frequently referenced by names under the .EU domain. There is no entity in charge of preserving the general content published under the .EU domain. The strategy adopted by memory institutions has been to preserve the web through the delegation of the responsibility to each national institution which leaves the content published under the .EU domain orphan regarding its preservation. Nonetheless, the Historical Archives of the European Union (HAEU), in cooperation with the EU Inter-institutional Web Preservation Working Group coordinated by the EU Office of Publications, has started a web archiving pilot project in late 2013 concerning the websites of EU institutions and bodies. They performed four complete crawls of 19 EU Institutional and Bodies websites in 2014 and extended this to include 50 EU Agencies in 2015 [19]. Arquivo.pt performed a first exploratory crawl of the .EU domain to gain insight into the preservation of the content published under this domain [23]. The initial idea was that the "brute-force" approach of preserving the .EU websites in general would also include most R&D projects websites hosted on this domain. However, the obtained results showed that this approach was too demanding for the resources we had available. Therefore, we decided to adopt a more selective approach. By combining open data sets and free search services, we have established a pragmatic framework that enables the automatic identification and preservation of R&D project URLs in contexts with very limited resources available.

3. EPHEMERA OF R&D WEBSITES

Everyday, more information is published on the web, from a simple blog post opinion to a research project funded by the European Union. However, the web is ephemeral. Only 20% of web pages remain unchanged after one year, which points towards a massive loss of information [37]. We performed an experiment to measure the ephemera of research websites funded by the European Union work programmes from FP4 (1994-1998) to FP7 (2007-2013). On the 27th November 2015, we tested the available projects URLs for each funding work programme (FP4 [4], FP5 [5], FP6 [6] and FP7 [7]), checking how many still referenced relevant content. The datasets containing the projects URLs was obtained from the European Union Open Data Portal datasets [8]. A comparison was made using the *title* on the datasets and the project URL content to test if each project URL was still referencing relevant content. The relevance criterion applied was that if at least half the words with 4 or more characters presented on the *title* were found on the content referenced by the project URL, the content was considered to be relevant. This method was applied on all work programmes with exception of FP7 that was humanly validated to build the test collection described in Section 4.

Table 1: Project URLs from the CORDIS dataset referencing relevant content distributed per work programme validated in 27 November, 2015.

	Nr. project URLs	Nr. project URLs with relevant content	% project URLS relevant content
FP4 (1994-1998)	853	58	7%
FP5 (1998-2002)	2 717	322	12%
FP6 (2002-2006)	2 401	715	30%
FP7 (2007-2013)	2 092	1 370	65%

The results presented on Table 1 show that 65% of the URLs of R&D projects funded by FP7 program were still available and referenced relevant content. A counterexample of a R&D project URL, presented on the FP7 dataset, that now references irrelevant content is www.oysterecover.eu. This URL is associated to the OYSTERECOVER project that studied scientific bases and technical procedures to recover the European oyster production, and now references a shopping website. The percentage of active and relevant project URLs decreased for older work programmes, reaching a percentage of only 7% for the FP4 work programme (1994-1998).

3.1 Preservation Coverage and Distribution

Our previous results showed that a significant percentage of project URLs is no longer available on the live-web and therefore its content may have been potentially lost forever. However, there are several web Archiving initiatives working to preserve the web as exhaustively as possible. Many of them focus on the preservation of each respective country web domain, with some exceptions like the US-based Internet Archive [13], a non-profit initiative that acts with a global scope.

We conducted an experiment to measure if the available project URLs referenced on the incomplete CORDIS datasets were preserved by web archives. For this purpose, we verified if at least one web-archived version of the referenced project URLs could be found by using the Time Travel Service [18, 21]. This service acts as gateway to query for archived versions of a web resource (Memento) across multiple publicly available web archives using the HTTP Memento Framework [26]. For each project URL, we queried the Time Travel Service for its *timemap* which provides a list of corresponding archived versions. If a project URL had an archived version between the time range of the corresponding work programme, we considered that the project URL had a valid archived version. The results of this experiment are presented on Table 2. It shows that 1 593 of the 2 092 FP7 project URLs have an archived version between 2007 and 2013, meaning that 76.1% of these projects URLs have an web-archived version. However, the amount of project URLs preserved decreases for the older work programmes, only 38.2% of the FP6 project URLs had a web-archived version, and 43.6% for FP4 project URLs.

Table 3 shows the distribution of the project URLs archived versions across web archives. For each project URL we counted how many web archives have a valid archived version. Most of the project URL archived versions are retrieved from web.archive.org, the time gate of the Internet Archive, with 76% preservation coverage of the FP7 project URLs followed by web.archive.bibalex.org with only 0.81% of the FP7 project URL preserved. This results show that EU-funded project URLs were mainly preserved by the US-based Internet Archive.

Table 2: Projects URLs on EU CORDIS datasets with a web-archived version.

	Nr. project URLs	Nr. project URLs with an archived version	% project URLs with an archived version
FP4 (1994-1998)	853	372	43.6%
FP5 (1998-2002)	2 717	1 661	61.1%
FP6 (2002-2006)	2 401	918	38.2%
FP7 (2007-2013)	2 092	1 593	76.1%

Table 3: Distribution of projects URL archived versions per web archive.

Time Gates	% FP4	% FP5	% FP6	% FP7	% Average
web.archive.org [13]	43.61	60.91	37.90	76.0	54.61
web.archive.bibalex.org [12]	12.54	22.56	21.90	0.72	14.43
webarchive.loc.gov [14]	0	1.80	0.58	0.43	2.81
webarchive.nationalarchives.gov.uk [46]	0.12	0.22	0	0.57	0.91
arquivo.pt [2]	0.47	0.55	0.24	0.67	0.48
wayback.archive-it.org [1]	0	0.04	0	0.81	0.21
wayback.vefsafn.is [36]	0.35	0.03	0.08	0.23	0.17
webarchive.parliament.uk [46]	0	0	0	0.19	0.05
www.webarchive.org.uk [24]	0	0	0	0.19	0.05
www.padi.cat [47]	0	0	0	0.04	0.01
collection.europarchive.org [32]	0	0	0	0.04	0.01

4. EVALUATION METHODOLOGY

This section describes the evaluation methodology used to compare the performance of the heuristics tested to automatically identify R&D projects websites. We present here test collection developed as well as the relevance criterion adopted.

4.1 Test Collection

A ground-truth is required to evaluate the performance of the proposed heuristics. We developed a test collection based on the FP7 dataset [7]. The objective was to build a list of carefully validated pairs of projects and corresponding project URLs. The CORDIS dataset contains several information fields about each funded project such as *acronym* (project acronym), *title* (description of the project) and *projectUrl* (URL for the project site or page). However, for most projects the URL was missing. Thus, we removed all the projects that had the *projectUrl* field blank, ending up with a list of 2 092 entries with *projectUrl* filled. Then, the following data cleansing steps were applied to the dataset:

1. Removed non-existent URLs or invalid URLs (return codes that are not 200s or 300s);
2. Followed all redirects and updated the *projectUrl* field with the target URL;
3. Removed non alphanumeric characters from the title fields;
4. Left and right trim each column and removal of multiple white spaces.

The dataset resulted in a list of 1 596 entries. However, this list was still not ready to be used as a test collection because there were project URLs referencing online content no longer related to the project. For example, some URLs projects referenced registrar sites, shopping sites or Chinese sites that became the new owners of the the domain names. A human validation was performed to overcome this situation, deleting entries where the *project URLs* were no longer related to the project. With this manual validation the test collection ended up with a list of 1 370 project entries with valid project URLs.

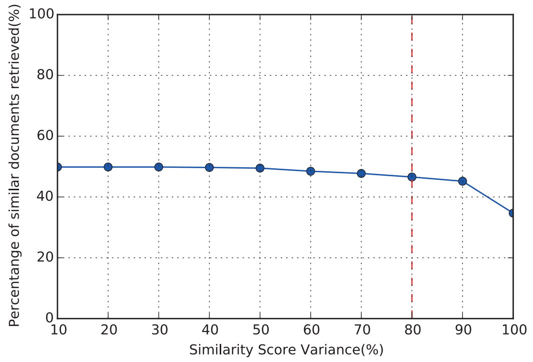


Figure 1: Fuzzy hashing threshold applied to identify relevant project URLs.

The search engines API have some limitations regarding how many queries can be made. For example, Google Custom Search Engine has a limitation of 100 search queries per day, and Bing Web Search API has a limitation of 5 000 queries a month for free usage. These limitations slowed the heuristics evaluation experiments to identify the R&D projects URLs. For our test collection of 1 370 entries, it would only be possible to experiment 3 heuristics each month. To be able to test several heuristics in a reasonable time, a smaller collection was built from the full test collection. This smaller test collection comprised a random sample of 300 entries from the base test collection, with a confidence level of 95% and a 5% margin of error [17].

4.2 Relevance Criterion

Ideally, the project URLs identified through an heuristic should match the project URL on the test collection. However, a strict string comparison to match URLs raises problems. For instance, it would not detect URLs with different domains but the same content like www.lipididiet.progressima.eu/ and lipididiet.eu/, nor the absence or presence of *www* hostname, www.hleg.de and hleg.de. Another problematic situation would be different paths names to the same content such as www.tacmon.eu/new/ and www.tacmon.eu/. Thus, we adopted an automatic content comparison approach by using hashing techniques instead of URL comparisons. However, the use of strict hashing techniques like MD5 [40] or SHA-1 [27] to verify if the content referenced by the project URL is relevant also present limitations. Project URLs that reference hidden dynamic content, for instance a simple blank space or a hidden HTML section inserted dynamically would result in totally different hash codes, leading wrongly to the conclusion that the content is not relevant. For this reason, we decided to apply a fuzzy hashing technique [34]. This technique allows us to overcome the previous problems since it generates an hash code proportional to the level of difference between contents. Noteworthy, the similarity threshold cannot be too high (e.g. 100%) because it would suffer from the limitations of strict hashing techniques causing the exclusion of relevant project URLs. On the other hand, the threshold cannot be too low because it would include irrelevant results. The similarity threshold was determined by gradually decreasing the similarity threshold and counting the percentage of relevant results retrieved. Figure 1 shows the percentage of relevant project URLs identified as the fuzzy hashing threshold value increased. We adopted a threshold of 80% for the matching score because the number of similar documents retrieved did not significantly varied below this value and a high percentage of similarity was found. Therefore, we defined that a project URL provided by a search engine is a relevant result for a given project if its content matches the content of the project URL defined on the test collection with a similarity level of at least 80%.

For each heuristic it was measured how well it performed on retrieving the project URL for each project entry on the test collection. An example of a relevant retrieval is when we apply a heuristic to query a search engine about a given project and it returns the URL of the home page of the website with a similarity score of more than 80% in comparison to the test collection project URL content.

5. HEURISTICS TO AUTOMATICALLY IDENTIFY R&D PROJECT URLS

Several heuristics to automatically identify project URLs on the live-web were tested. The main idea of these heuristics is to use search engines retrieval capabilities to identify URLS of research project websites.

5.1 +Acronym +Title

This heuristic consists on querying the search engines using the Acronym and Title fields of the FP7 dataset, despite its name provides a textual description of the project. An example of a query submitted to a search engine using this heuristics is: "IMPACT Impact Measurement and Performance Analysis of CSR".

5.2 +Acronym +Title -Cordis

This heuristic consists on querying the search engine using the Acronym and Title fields but excluding the results from site cordis.europa.eu. The rational behind this exclusion is that search engine results can be biased towards results hosted on the CORDIS site since the query terms used were obtained from the CORDIS datasets. An example of a query submitted to a search engine using this heuristics is: "IMPACT Impact Measurement and Performance Analysis of CSR -site:cordis.europa.eu".

5.3 +Acronym +Title -Cordis -EC

This heuristic is the same as the **+Acronym +Title -Cordis** but also excludes the site ec.europa.eu. An example of a query submitted to a search engine using this heuristics is: "IMPACT Impact Measurement and Performance Analysis of CSR -site:cordis.europa.eu -site:ec.europa.eu".

5.4 +Acronym +Title -Cordis -EC +CommonTerms

This heuristic aims at improving the results returned by search engines through the inclusion of additional query terms commonly used on the content referenced by existing project URLs. The most frequent words extracted from the test collection projects websites content, were identified and then the queries were built by adding these common terms to the query issued to the search engine.

$$\mathbf{v_m} = sort\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_{d_i}\right) \qquad (1)$$

The method to compute these terms was established through a bag of words model, generating a features vector for each project site corpus $\{\mathbf{v}_{d_1},\ldots,\mathbf{v}_{d_n}\}$, where each feature represents a word weighted by a TF-IDF weighting scheme [38]. Then, the mean of all features vectors sorted by the highest weighted features was calculated (Equation 1). Table 4 present the top 10 features retrieved $\{\mathbf{v}_{m_1},\ldots,\mathbf{v}_{m_{10}}\}$. That is, the top 10 most common terms in the text of project URLs after removing irrelevant words such as stopwords. An example of a query derived using this heuristic is: "IMPACT Impact Measurement and Performance Analysis of CSR **project** -site:cordis.europa.eu -site:ec.europa.eu".

Table 4: Top 10 most common terms in the web content referenced by project URLs validated to build the test collection.

Position	Term	Average TF-IDF
1	project	0.048
2	research	0.023
3	european	0.021
4	home	0.017
5	news	0.017
6	eu	0.015
7	new	0.014
8	2015	0.014
9	read	0.014
10	partners	0.014

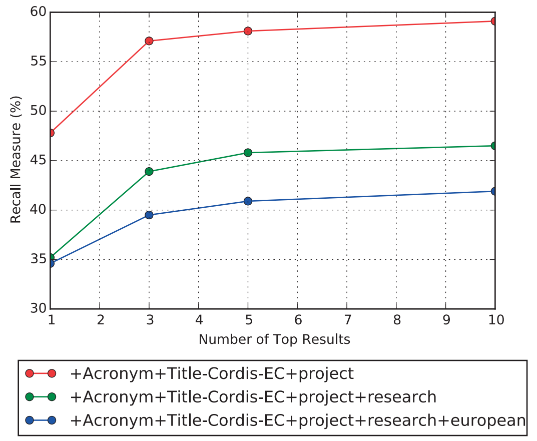


Figure 2: Recall of heuristics using additional common terms in queries.

6. HEURISTICS TUNING AND PERFORMANCE

Each heuristic performance was measured and compared through *recall* (2), *precision* (3) and *f-measure* (4) metrics to evaluate the success of the proposed heuristic on identifying the project URLs of the test collection. The scores were measured by analyzing the Top 1, Top 3, Top 5 and Top 10 results obtained through the Bing Web Search API.

$$recall = \frac{|\{\text{relevant documents}\}\bigcap\{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|} \qquad (2)$$

$$precision = \frac{|\{\text{relevant documents}\}\bigcap\{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|} \qquad (3)$$

$$f\text{-}measure = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \qquad (4)$$

Before comparing the performance between the described heuristics, the selection of common terms added to the heuristic **+Acronym +Title -Cordis -EC +CommonTerms** was tuned looking for the combination with the highest potential to provide the best performance. The following term combinations were tested: {*project*}, {*project*,*research*}, {*project*,*research*,*european*}. Based on the results presented on Figure 2, it was determined that the usage of only one term {*project*} provided the best results. Increasing the number of terms restricts too much the query scope obtaining lower recall values. Therefore, we decided to adopt only the addtitional common term *project* for the heuristic **+Acronym +Title -Cordis -EC +CommonTerms** and named it **+Acronym +Title -Cordis -EC +project**.

Table 5: Recall of each heuristic when identifying project URLs on the live-web.

	Top 1	Top 3	Top 5	Top 10
+Acronym +Title	44.0%	56.3%	*64.0%*	*66.0%*
+Acronym +Title -Cordis	44.9%	55.1%	58.1%	60.1%
+Acronym +Title -Cordis -EC	46.8%	56.1%	58.5%	60.5%
+Acronym +Title -Cordis -EC +project	*47.8%*	*57.1%*	58.1%	59.1%

Table 6: Precision of each heuristic.

	Top 1	Top 3	Top 5	Top 10
+Acronym +Title	44.0%	*18.8%*	*12.8%*	6.6%
+Acronym +Title -Cordis	44.9%	18.4%	11.6%	6.0%
+Acronym +Title -Cordis -EC	46.8%	18.7%	11.7%	6.0%
+Acronym +Title -Cordis -EC +project	*47.8%*	*19.0%*	11.6%	5.9%

Table 5 presents the score results for *recall* obtained for all the heuristics. As expected, it shows that increasing the number of top results retrieved increases the *recall* score. The heuristic with best recall (47.8%) at the TOP 1 results is the **+Acronym +Title +project -Cordis -EC**, but this is the worst heuristic at the Top 10 results (59.1% against **+Acronym +Title** 66%). Since this heuristic query contains more terms, it is more specific, becoming more precise at the Top 1 results, but the lack of generalization makes it worse with more results returned. Therefore, we conclude that is the most suitable heuristic when we aim to achieve more precise identification and retrieval of project URLs. The **+Acronym +Title** heuristic is the more general query and so it returns more results. It is most suitable when the objective is to obtain the highest coverage of project URLs to be preserved without limiting resources and preserving also some less relevant sites.

Table 6 indicates the *precision* scores obtained. As expected, they decrease as more results returned are considered because each query has only 1 valid result identified on the test collection. The heuristic that presented higher precision values was **+Acronym +Title -Cordis -EC +project** with 47.8%.

The *F-measure* metric provides a combination of the *recall* and *precision* values. Those results are presented on Table 7 and show that **+Acronym +Title +project -Cordis -EC** has the highest score with 47.8% at Top 1.

7. A FIRST SELECTIVE CRAWL TO PRESERVE R&D WEBSITES

The experiments previously described were also tested using Google Custom Search Engine. This provided better results with an overall

Table 7: f-measure of each heuristic.

	Top 1	Top 3	Top 5	Top 10
+Acronym +Title	44.0%	28.2%	*21.3%*	*12.0%*
+Acronym +Title -Cordis	44.9%	27.6%	19.3%	10.9%
+Acronym +Title -Cordis -EC	46.8%	*28.1%*	19.5%	10.9%
+Acronym +Title -Cordis -EC +project	*47.8%*	28.5%	19.3%	10.7%

Table 8: Data related to R&D project websites collected by the crawler for preservation.

Nr. project URL seeds	20 429
Nr. web files crawled	10 449 947
Nr. hosts crawled	72 077
Stored content size (compressed)	1.4 TB

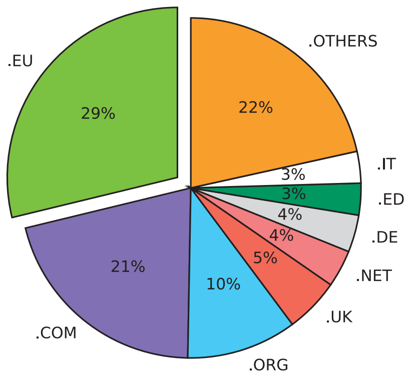


Figure 3: Retrieved R&D projects websites domain distribution.

recall gain of 5% against Bing, but the limitation of 100 queries per day made it impracticable because the testing procedure of the heuristics was too slow. We believe that the ability to do 5 000 queries/month of Bing Web Search API compensate for the slightly worse performance. For that reason Bing was the search engine that we used for the identification of new project URLs for R&D websites. The obtained results showed that the heuristic **+Acronym +Title -Cordis -EC +CommonTerms** achieved the best performance recovering project R&D URLs using the first result (Top 1), so it was the elected heuristic to apply to the incomplete FP7 projects dataset that presented 23 588 missing project URLs. The following work flow was executed to identify and preserve R&D project websites using the heuristics developed:

1. Extracted all project entries where *project URL* field was missing from the FP7 dataset;

2. Executed the heuristic **+Acronym +Title -Cordis -EC +CommonTerms** on FP7 projects dataset to recover missing URLs;

3. Used the newly identified URLs has seeds to the Heritrix crawler [35];

4. Harvested these project URLs and preserved this information.

After applying this workflow to the FP7 dataset, we identified 20 429 new URLs from the 23 588 entries with missing project URLs. That is, we improved the completeness of the CORDIS dataset by 86.6% regarding the project URLs information. About 3 159 entries did not return any URL, most probably because the project site does not exist any more, or never did.

These 20 429 new project URLs were used as seeds to a new selective crawl that resulted on the collection of 10 449 947 web files, fulfilling a total of 1.4 TB of information compressed on ARC files as presented on Table 8. This selective crawl was configured to crawl all mime types, following links until 5 hops from the project URL seed, with a limitation of 10 000 files per site.

Figure 3 depicts the project URLs domain distribution on the crawl. Most of the crawled R&D project sites were hosted under the .EU domain. So, we measured the overlap between the preserved content using the **+Acronym +Title -Cordis -EC +CommonTerms** heuristic and the crawled content obtained from our previous .EU domain crawl [23]. Using the OpenSearch [15] API available at arquivo.pt/resawdev, we queried if the projects URLs obtained had been previously harvested. Only 9% of the retrieved R&D projects websites were previously preserved by the .EU crawl.

8. CONCLUSIONS

Research & Development (R&D) projects rely on their websites to publish valuable information about their activities and achievements. However, these websites frequently disappear after the project funding ends. The European Union Open Data Portal provides information about R&D projects funded by the European Union. We tested the available projects URLs for each funding work programme. The obtained results showed that 65% of the URLs of R&D projects funded by FP7 program (2007-2013) were still valid. However, the percentage of valid project URLs decreased for older work programmes, reaching a percentage of only 7% for the FP4 work programme (1994-1998). The obtained results also showed that 76.1% of these projects URLs had an web-archived version. However, the amount of project URLs preserved decreased for the older work programmes. Only 43.6% of the FP4 project URLs had a web-archived version. The results also showed that EU-funded project URLs were mainly preserved by the US-based Internet Archive.

The main objective of this work was to study and develop an automatic mechanism that enables the identification of R&D project URLs of websites to be preserved without requiring strong human intervention or demanding computer resources. We designed and analyzed several heuristics that aimed to automatically identify missing project URLs combining live-web search engines and information publicly available about the projects. The experimental results showed that the most precise heuristic was being able to retrieve 47.8% of the missing projects URLs. This heuristic was applied to identify and recover the 23 588 project URL missing on the CORDIS dataset about the FP7 work programme. It successfully retrieved 20 429 URLs with high potential of being the original project URL or related content (86.6%).

The newly identified project URLs were used as seeds to a selective crawl aimed to preserve EU-funded R&D project websites. 10 449 947 web files were crawled from 72 077 hosts, fulfilling a total of 1.4 *Terabytes* of information compressed on ARC files. Most of the crawled R&D project sites were hosted under the .EU domain. Only 9% of the retrieved R&D projects websites were previously preserved by the .EU crawl performed by the Arquivo.pt web archive in 2015. These R&D project websites content may have changed during their lifetime, and this information is irrecoverable unless a web archive holds past versions of these sites.

As societies evolve and become more aware of the importance of preserving born-digital content, it is expectable that R&D project websites will become systematically identified, archived and preserved during administrative work flows. If so, the described heuristics will become necessary only for exceptional situations. Meanwhile, automatic heuristics are crucial to preserve online scientific outputs.

9. FUTURE WORK

In future work these heuristics could be improved to reach higher levels of *recall*. One way to try to improve these heuristics is to include more research network and funding websites that were not previously identified, such as erc.europa.eu. Applying search operators to restrict results to HTML content could also enhance the overall *recall* and contribute to higher quality project URLs seeds. URLs that reference content on other formats (e.g. PDF) are less likely to reference the home page of the project websites. Other methodologies and term combinations to extract describing words and improve query results could also be studied.

The test collection could be extended with additional results such as several relevant project URLs per each project entry. These extensions would accommodate situations such as projects that adopted different URLs across time or that provide several versions of the project URL in different languages.

Machine Learning algorithms can be trained using the test collection built during this study to automatically classify if an identified project URL is actually a R&D project website. The application of these algorithms on the websites retrieved by the studied heuristics could reduce the number of false positive seeds added to the crawler.

10. ACKNOWLEDGMENTS

We would like to acknowledge Cátia Laranjeira for her precious proof-reading of this article.

11. REFERENCES

[1] Archive-It - Web Archiving Services for Libraries and Archives. https://archive-it.org/.

[2] Arquivo.pt: pesquisa sobre o passado. http://arquivo.pt/.

[3] Bing Search API Web | Microsoft Azure Marketplace. http://datamarket.azure.com/dataset/bing/searchweb.

[4] CORDIS - EU research projects under FP4 (1994-1998) Datasets. https://open-data.europa.eu/en/data/dataset/cordisfp4projects.

[5] CORDIS - EU research projects under FP5 (1998-2002) Datasets. https://open-data.europa.eu/en/data/dataset/cordisfp5projects.

[6] CORDIS - EU research projects under FP6 (2002-2006) -Datasets. https://open-data.europa.eu/en/data/dataset/cordisfp6projects.

[7] CORDIS - EU research projects under FP7 (2007-2013) Datasets. http://open-data.europa.eu/en/data/dataset/cordisfp7projects.

[8] European Union Open Data Portal. http://open-data.europa.eu/en/data/.

[9] FAROO - Free Search API. http://www.faroo.com/hp/api/api.html.

[10] FCT - Fundação para a Ciência e a Tecnologia. http://www.fct.pt/index.phtml.en.

[11] Google Custom Search Engine. https://cse.google.com/.

[12] International School of Information Science (ISIS). http://www.bibalex.org/isis/frontend/home/home.aspx.

[13] Internet Archive: Digital Library of Free Books, Movies, Music & Wayback Machine. https://archive.org/index.php.

[14] Library of Congress. https://www.loc.gov/.

[15] OpenSearch. http://www.opensearch.org/Home.

[16] Research resources and outputs. https://github.com/arquivo/Research-Websites-Preservation.

[17] Test collection 300 samples. https://github.com/arquivo/Research-Websites-Preservation/blob/master/datasets/fp7-golden-dataset-300.csv.

[18] Time Travel. http://timetravel.mementoweb.org/.

[19] Websites Archives of EU Institutions. http://www.eui.eu/Research/HistoricalArchivesOfEU/WebsitesArchivesofEUInstitutions.aspx.

[20] S. G. Ainsworth, A. AlSum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How Much of the Web Is Archived? pages 1–10, 2012.

[21] S. Alam, M. L. Nelson, H. Van de Sompel, L. L. Balakireva, H. Shankar, and D. S. H. Rosenthal. Web Archive Profiling Through CDX Summarization. *Research and Advanced Technology for Digital Libraries*, 9316:3–14, 2015.

[22] ARCOMEM. Arcomem. https://web.archive.org/web/20130426060455/http://www.arcomem.eu/, October 2011.

[23] D. Bicho and D. Gomes. A first attempt to archive the .EU domain Technical report. http://arquivo.pt/crawlreport/Crawling_Domain_EU.pdf, 2015.

[24] British Library. UK Web Archive. http://www.webarchive.org.uk/ukwa/, 2011.

[25] S. Chakrabarti, M. Van Den Berg, and B. Dom. Focused crawling: A new approach to topic-specific Web resource discovery. *Computer Networks*, 31(11):1623–1640, 1999.

[26] H. V. de Sompel, M. Nelson, and R. Sanderson. Http framework for time-based access to resource states – memento. RFC 7089, RFC Editor, December 2013.

[27] D. Eastlake and P. Jones. Us secure hash algorithm 1 (sha1). RFC 3174, RFC Editor, September 2001. http://www.rfc-editor.org/rfc/rfc3174.txt.

[28] E. Evangelou, T. A. Trikalinos, and J. P. Ioannidis. Unavailability of online supplementary scientific information from articles published in major journals. *The FASEB Journal*, 19(14):1943–1944, 2005.

[29] M. Faheem and P. Senellart. Intelligent and adaptive crawling of web applications for web archiving. In *Web Engineering*, pages 306–322. Springer, 2013.

[30] D. H.-L. Goh and P. K. Ng. Link decay in leading information science journals. *Journal of the American Society for Information Science and Technology*, 58(1):15–24, 2007.

[31] D. Gomes and M. J. Silva. Modelling information persistence on the web. In *ICWE '06: Proceedings of the 6th international conference on Web engineering*, pages 193–200, New York, NY, USA, 2006. ACM Press.

[32] Internet Memory Foundation. Internet Memory Foundation. http://internetmemory.org/en/.

[33] M. Klein and M. L. Nelson. Evaluating methods to rediscover missing web pages from the web infrastructure. In *Proceedings of the 10th annual joint conference on Digital libraries*, pages 59–68. ACM, 2010.

[34] J. Kornblum. Identifying almost identical files using context triggered piecewise hashing. *Digital Investigation*, 3(SUPPL.):91–97, 2006.

[35] G. Mohr, M. Stack, I. Ranitovic, D. Avery, and M. Kimpton. An introduction to Heritrix: An open source archval quality Web crawler. In *4th International Web Archiving Workshop*, number 2004, 2004.

[36] National and University Library of Iceland. Vefsafn - English. http://vefsafn.is/index.php?page=english, 2011.

[37] A. Ntoulas, J. Cho, and C. Olston. What's new on the web?: the evolution of the web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web*, pages 1–12. ACM Press, 2004.

[38] J. Ramos, J. Eden, and R. Edu. Using TF-IDF to Determine Word Relevance in Document Queries. *Processing*, 2003.

[39] T. Risse, S. Dietze, W. Peters, K. Doka, Y. Stavrakas, and P. Senellart. Exploiting the social and semantic web for guided web archiving. In *Theory and Practice of Digital Libraries*, pages 426–432. Springer, 2012.

[40] R. L. Rivest. The md5 message-digest algorithm. RFC 1321, RFC Editor, April 1992. http://www.rfc-editor.org/rfc/rfc1321.txt.

[41] B. Sampath Kumar and K. Manoj Kumar. Decay and half-life period of online citations cited in open access journals. *The International Information & Library Review*, 44(4):202–211, 2012.

[42] J. L. Shipman, M. Klein, and M. L. Nelson. Using web page titles to rediscover lost web pages. *arXiv preprint arXiv:1002.2439*, 2010.

[43] A. S. Sife and R. Bernard. Persistence and decay of web citations used in theses and dissertations available at the sokoine national agricultural library, tanzania. *International Journal of Education and Development using Information and Communication Technology*, 9(2):85, 2013.

[44] D. Spinellis. The decay and failures of web references. *Communications of the ACM*, 46(1):71–77, 2003.

[45] O. Tajeddini, A. Azimi, A. Sadatmoosavi, and H. Sharif-Moghaddam. Death of web citations: a serious alarm for authors. *Malaysian Journal of Library & Information Science*, 16(3):17–29, 2011.

[46] The National Archives. UK Government Web Archive | The National Archives. http://www.nationalarchives.gov.uk/webarchive/, 2011.

[47] The Web Archive of Catalonia. The Web Archive of Catalonia. http://www.padi.cat/en.

POSTERS //

A Demonstration of BitCurator Access Webtools and Disk Image Redaction Tools

Christopher A. Lee and Kam Woods
School of Information and Library Science
University of North Carolina
216 Lenoir Drive, CB #3360
1-(919)-966-3598
callee@ils.unc.edu; kamwoods@email.unc.edu

ABSTRACT

BitCurator Access is developing open-source software that supports the provision of access to disk images through three exploratory approaches: (1) building tools to support web-based services, (2) enabling the export of file systems and associated metadata, (3) and the use of emulation environments. This demonstration will highlight two BitCurator Access software products: BitCurator Access Webtools which supports browser-based search and navigation over data from disk images, and a set of scripts to redact sensitive data from disk images.

Categories and Subject Descriptors

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

General Terms

Provenance; Data Triage; Digital Forensics.

Keywords

Digital forensics; preservation; DFXML; metadata; privacy; collections; web access; redaction

1. BITCURATOR ACCESS PROJECT

The BitCurator Access project began on October 1, 2014 and will end on September 30, 2016. Funded through a grant from the Andrew W. Mellon Foundation, BitCurator Access is developing open-source software that supports the provision of access to disk images through three exploratory approaches: (1) building tools to support web-based services, (2) enabling the export of file systems and associated metadata, (3) and the use of emulation environments. Also closely associated with these access goals is redaction. BitCurator Access is developing tools to redact files, file system metadata, and targeted bitstreams within disks or directories.

BitCurator Access focuses on approaches that simplify access to raw and forensically-packaged disk images; allowing collecting institutions to provide access environments that reflect as closely as possible the original order and environmental context of these materials. The use of forensic technologies allows for detailed metadata to be generated to reflect the provenance of the materials, the exact nature of the file-level items they contain, and the metadata associated with both file-level items and data not observed within the file system (but still accessible within the original materials). We are freely disseminating the BitCurator Access software products under an open source (GPL, Version 3) license. All existing software upon which the products are built is also either open-source or public domain.

This demonstration will highlight two BitCurator Access software products: BitCurator Access Webtools which supports browser-based search and navigation over data from disk images, and a set of scripts to redact sensitive data from disk images. We have previously reported on support for workflows that employ BCA Webtools and Emulation-as-a-Service (EaaS) [3].

2. BITCURATOR ACCESS WEBTOOLS

The BitCurator Access project has developed BCA Webtools, which is a suite of software (based on an earlier prototype called DIMAC [2]) that allows users to browse a wide range of file systems contained within disk images using a web browser. It is intended to support access requirements in libraries, archives, and museums preserving born-digital materials extracted from source media as raw or forensically-packaged disk images.

BCA Webtools uses open source libraries and toolkits including The Sleuth Kit, PyTSK, and the Flask web microservices framework. It uses PyLucene along with format-specific text-extraction tools to index the contents of files contained in disk images, allowing users to search for relevant content without individually inspecting files. BCA Webtools is distributed with a simple build script that deploys it as a Vagrant virtual machine running the web service.

The application can parse raw and E01-packaged images containing FAT16, FAT32, NTFS, HFS+, and EXT 2/3/4 file systems, and allows users to navigate the file system contents, download individual files, and search the contents within a simple web interface.

3. REDACTION TOOLS

Digital media acquisitions in libraries, archives and museums often contain data that may be classified as private, sensitive, or individually identifying, and the complexity and volume of information being collected demands automation to ensure that risks of inadvertent disclosure are minimized.

Currently, there are relatively few open source redaction tools capable of addressing these needs. BitCurator Access is target specific areas of software development, including:

- Redacting specific bitstreams from raw disk images
- Creating redacted copies of forensically-packaged disk images
- Building redaction overlays that can applied to disk images in an access context, masking out specific files and directories
- Redacting metadata from commonly used file formats, including Office and PDF files.

This demonstration will include modifications to and adaptations of existing Digital Forensics XML tools [1] that provide support for the above activities. Specifically, we will demonstrate a Python tool for redacting sequences of data from disk images matching one or more pattern(s) provided as arguments on the command line or in a configuration file.

The demonstrated redaction tool that is neither file system nor file format sensitive by default, although it may operate using the output of tools that output file system statistics including byte runs associated with individual files and directories identified within recognized file systems. The tool will also perform redaction operations on relevant byte sequences identified in raw data streams, whether or not they are presented in the form of disk images.

4. ACKNOWLEDGMENTS

Development of the BitCurator environment and BitCurator Access Webtools have been supported by the Andrew W. Mellon Foundation. The BitCurator Access team is composed of Alex Chassanoff, Christopher Lee, Sunitha Misra, and Kam Woods. Greg Jansen contributed significantly to the development and documentation of the redaction tools featured in this demonstration.

5. REFERENCES

[1] Garfinkel, S. Digital Forensics XML and the DFXML Toolset. *Digital Investigation* 8 (2012), 161-174.

[2] Misra, S., Lee, C. A., and Woods, K. 2014. A Web Service for File-Level Access to Disk Images. *Code4Lib Journal* 25 (2014).

[3] Woods, K., Lee, C. A., Stobbe, O., Liebetraut, T., and Rechert, K. 2015. Functional Access to Forensic Disk Images in a Web Service. In *Proceedings of iPRES 2015*. University of North Carolina, Chapel Hill, NC, 191-195.

Assigning Rights Statements to Legacy Digital Collections

Laura Capell
University of Miami Libraries
1300 Memorial Drive
Coral Gables, FL 33124
1-305-284-9328
l.capell@miami.edu

Elliot Williams
University of Miami Libraries
1300 Memorial Drive
Coral Gables, FL 33124
1-305-284-4730
edwilliams@miami.edu

ABSTRACT

This poster reports on a project at the University of Miami Libraries to evaluate the rights status of legacy materials that have been digitized for online access in the UM Digital Collections, and to assign item-level rights statements to over 52,000 items.

1. BACKGROUND

The University of Miami Libraries began a project in the fall of 2015 to evaluate the rights status of legacy materials digitized for online access in the UM Digital Collections. The project objective is to categorize the contents of our digital collections based on the parameters established by RightsStatments.org [1].

The Libraries' Cuban Heritage Collection, Special Collections, and University Archives contain a wealth of resources documenting the history and culture of the Caribbean basin, with a focus on Florida, Cuba, and Haiti. Over the past fifteen years, thousands of items from these collections have been digitized to facilitate online access, including publications, photographs, manuscripts, architectural drawings, maps, oral histories, and audio and video recordings. In addition to the wide variety of formats and geographical locations represented in the digital collections, they also span a large timeframe, from the 16th century to the present. This diversity is beneficial for researchers, but it presents challenges for creating accurate rights statements.

At the start of the project, the majority of the Libraries' digital collections contained little to no rights-related information in their metadata. While rights status at the collection level was often discussed during the project planning stage, specific rights information was not included in the item-level metadata unless the Libraries explicitly received permission to digitize from the copyright holder. Often, the exact rights status was not known, with many materials falling into the gray area of orphan works.

However, as we ramp up outreach efforts to engage researchers in traditional and nontraditional uses of our digital collections, we want to empower our users to make better-informed decisions about potential uses of our online resources. Therefore, we decided to conduct a systematic review of our digitized content to determine the rights status and provide appropriate rights information in the item-level metadata.

This project also coincides with plans to create a Florida service hub for the Digital Public Library of America (DPLA), which would provide the Libraries a pathway to contribute our content to DPLA. The inclusion of rights metadata is a prerequisite for DPLA, so the timing of this project is perfect as we begin to assess potential metadata cleanup and transformations necessary to prepare for DPLA harvesting.

2. WORKFLOW

Our publicly accessible digital holdings are comprised of over 52,000 items spread over 120 distinct digital collections, and at the start of the project, less than 5,800 items had any specific rights information in the metadata. Our initial plan was to conduct a collection-level rights assessment for each digital collection, but we quickly realized that the content within each collection often contains a multitude of different rights scenarios. This is especially true for manuscript collections, which can include materials by numerous creators spanning a wide date range, with some content in the public domain but much still falling under copyright. Many items lack definitive identifying information, such as the creator or date of creation, making it challenging to determine the rights status. In order to achieve a higher level of accuracy in our assessment, we decided to review and assign rights categories at the item level.

The first step was to review relevant deeds of gift to better understand the rights landscape for each collection. We were able to note when the donor retained rights to their materials and when they had transferred those rights to the University of Miami. We also noted collections that were either purchased, had no deed of gift, or lacked any substantive rights information. Although it did not always provide definitive answers, this step did enrich our contextual understanding of the collections.

Next, we assessed each collection, using item-level metadata exported from CONTENTdm. To enable the project to move forward quickly, we split the work in half, with each of us separately reviewing metadata for a collection and assigning rights statuses. We met frequently to go over questions that arose, researching and discussing the more challenging scenarios we uncovered. We have documented the rationale behind our decisions at the collection level to provide context in case future reevaluations are needed.

We created a decision matrix to ensure consistency during the evaluation process. The matrix addresses the most common rights scenarios we have encountered for published and unpublished materials with personal, corporate, or government authors. It also accounts for the country of creation, since a large percentage of our materials originated in Cuba, which entails different copyright considerations. The matrix is a fluid document that has evolved over time as we encounter new rights scenarios, but it has been an invaluable tool to simplify decision making and remove as much guesswork as possible from the evaluation process.

After assessing the collections, we added two rights-related fields to our Dublin Core metadata records in CONTENTdm. The first field is a local rights statement, which includes any known information about the copyright holder and a link to our digital collections copyright webpage. The second field contains the RightsStatements.org label and URI. This allows us to provide both customized local rights information and a standardized, machine-actionable rights statement as recommended by RightsStatements.org [2]. (See Table 1 below.)

3. CHALLENGES

Our determinations are based on the information available in the metadata, and we do not have time to conduct in-depth research on thousands of items. Therefore, the status we assign is our best guess based on the information available, and if additional information comes to light in the future, we will update the rights status accordingly.

Over the course of the project, we have encountered several challenges in determining rights ownership for such a wide variety of materials. One of the primary challenges has been orphan works, especially undated, unpublished materials where little to nothing is known about the creator. Our hope was to assign a definitive rights status to every item, clearly identifying materials as being in copyright or in the public domain, but we encountered a large amount of unpublished material with no date or creator information. In these situations, we chose to label these items as "copyright undetermined" since they lack information to assign an accurate rights status.

We have also grappled with determining the extent to which the donor held copyright to the materials in the collection. For example, if a niece donated her deceased uncle's photography collection, did she inherit the intellectual rights to the images to be able to transfer the rights to the library? Often, there were few clear answers, but reviewing the donation terms in the deeds of gift did provide us with the background to better understand the provenance and context of the various collections.

An additional difficulty has been determining whether an item should be considered published or unpublished. Publication status is very important under U.S. copyright law, but the large variety of materials found in a modern manuscript collection can create questions about what counts as publication. Again, without examining individual items, it can be challenging to determine whether certain types of materials, such as early postcards or mimeographed flyers, were indeed published.

Another challenge has been deciphering international copyright issues. While our focus is to determine the legal status of materials in the United States, in some cases copyright may vary according to the country of origin. For the large amount of Cuban material in our collections, we have reviewed Cuban copyright legislation, including international treaty regimes and varying definitions of public domain. Unpublished personal and corporate materials from Cuba have proven to be especially challenging, because of nuances in Cuban copyright law that differ from U.S. law. Given the transnational nature of our materials, the recommendations made by Europeana and DPLA have been invaluable for helping frame our rights statements in an international context.

4. REFERENCES

[1] RightsStatements.org. Europeana and Digital Public Library of America.
http://rightsstatements.org/page/1.0/?language=en

[2] International Rights Statements Working Group. 2016. *Rightsstatements.org White Paper: Requirements for the Technical Infrastructure for Standardized International Rights Statements.* Europeana and Digital Public Library of America.
http://rightsstatements.org/files/160322requirements_for_the_technical_infrastructure_for_standardized_international_rights_statements_v1.1.pdf

Table 1. Local and Standardized Rights Statements Used in Dublin Core Metadata Records

Local Rights Statement	Standardized Rights Statement
This material is protected by copyright. Copyright is held by the creator.	In Copyright http://rightsstatements.org/vocab/InC/1.0/
This material is protected by copyright. Copyright is held by […].	In Copyright http://rightsstatements.org/vocab/InC/1.0/
This material is protected by copyright. Copyright was originally held by […], but was transferred to the University of Miami.	In Copyright http://rightsstatements.org/vocab/InC/1.0/
This material is protected by copyright. The copyright owner is unknown or unidentifiable.	In Copyright – Rights-holder(s) Unlocatable or Unidentifiable http://rightsstatements.org/vocab/InC-RUU/1.0/
This material is in the public domain in the United States.	No Copyright – United States http://rightsstatements.org/vocab/NoC-US/1.0/
The copyright and related rights status of this material is unknown.	Copyright Undetermined http://rightsstatements.org/page/UND/1.0/
No copyright or related rights are known to exist for this material, but conclusive facts may be missing or ambiguous.	No Known Copyright http://rightsstatements.org/vocab/NKC/1.0/
Copyright status as noted on the item: "[…]"	*Select the appropriate rights statement*

Autonomous Preservation Tools in Minimal Effort Ingest

Asger Askov Blekinge
State and University Library,
Denmark
+45 8946 2100
abr@statsbiblioteket.dk

Bolette Ammitzbøll Jurik
State and University Library,
Denmark
+45 8946 2322
baj@statsbiblioteket.dk

Thorbjørn Ravn
Andersen
State and University Library,
Denmark
+45 8946 2317
tra@statsbiblioteket.dk

ABSTRACT

This poster presents the concept of *Autonomous Preservation Tools*, as developed by the State and University Library, Denmark. The work expands the idea of *Minimal Effort Ingest*, where most preservation actions such as Quality Assurance and enrichment of the digital objects are performed after content is ingested for preservation, rather than before. We present our Newspaper Digitisation Project as a case-study of real-world implementations of Autonomous Preservation Tools.

Keywords

Long term Preservation, Object Repository, Minimal Effort Ingest, Autonomous Preservation Tools, Tree Iterator, Preservation Actions, Quality Assurance, OAIS

1. INTRODUCTION

The State and University Library, Denmark, would like to present an expansion of the Minimal Effort Ingest model [1, 2]. In Minimal Effort Ingest most of the preservation actions are postponed and handled within the repository, when resources are available. We suggest organising these actions using *Autonomous Preservation Tools* – similar to software agents – rather than a static workflow system. This adds flexibility to the repository, as it allows for easy updates, removal or addition of workflow steps. We present the Danish newspaper digitisation project as a case study of Autonomous Preservation Tools. Firstly we introduce the concepts of Minimal Effort Ingest (section 2) and Autonomous Preservation Tools (section 3). We then present the newspaper digitization project (section 4), and how these concepts have been implemented.

2. MINIMAL EFFORT INGEST

When ingesting a collection into an object repository for long term preservation it is common to follow the OAIS reference model [3]. In OAIS quality assurance (QA) and enrichments are performed on the submission information package (SIP) before this is ingested into the repository. The Minimal Effort Ingest [1, 2] idea builds on OAIS, but performs the QA and enrichment actions on the archival information package (AIP) inside the repository. The aim is to secure the incoming data quickly, even when resources are sparse.

3. AUTONOMOUS PRESERVATION TOOLS

In Minimal Effort Ingest preservation actions should not be orchestrated by a static ingest workflow, but rather be carried out *when resources are available*. From this concept, we developed the idea of Autonomous Preservation Tools.

We use the OAIS term Archive Information Package (AIP) to denote an object stored in a preservation system. We define Preservation Tools as tools that can be used on such an AIP. The implementation of such a tool is very dependent both on the preservation system and the format of the AIP. An AIP could be anything from a simple file to a complex container format or interlinked structure.

An Autonomous Preservation Tool is an extension of a normal preservation tool. Traditionally preservation tools are explicitly invoked on an AIP as part of a static workflow. Autonomous Preservation Tools can discover AIPs to process on their own.

We further assume that AIPs maintain an account of past events. In Digital Preservation such an account can be important for showing data authenticity and provenance, so many repository systems implement this already. From this account the Autonomous Preservation Tool can determine whether it has already processed an AIP or not.

The order in which the Autonomous Preservation Tools process an AIP can be important, such as whether checksums were generated before or after additional metadata was added. Thus, Autonomous Preservation Tools must be able to find the AIPs they have not yet processed, but which have already been processed by specific other Autonomous Preservation Tools.

With Autonomous Preservation Tools the need for a fixed workflow disappears and is replaced with a decentralised implicit workflow. Rather than defining a fixed sequence of steps an AIP must go through, you define the set of events that an AIP must have experienced. Each Autonomous Preservation Tool corresponds to one such event, which it sets when it processes the AIP. The workflow will be a number of Autonomous Preservation Tools each looking for AIPs to process, until each tool has processed every AIP.

This approach brings a great deal of flexibility:

- Removing an Autonomous Preservation Tool is a local operation. No other Autonomous Preservation Tools or workflows will be affected.
- When a new Autonomous Preservation Tool is added, it will automatically start processing old AIPs as well as new. No migration steps are needed.
- When an Autonomous Preservation Tool is changed, it can be marked as a new Autonomous Preservation Tool, and thus start processing all previously processed AIPs. Alternatively, the tool can be configured to continue where it left off, and thus only process new AIPs.

4. CASE STUDY

The Danish newspaper digitisation project [1] is an in-production example of Minimal Effort Ingest using Autonomous Preservation Tools. In this project we receive scanned newspaper pages in batches of about 25,000 pages along with MIX[2], MODS[3] and ALTO[4] metadata. We receive two batches a day and a total of about 30 million newspaper pages throughout the duration of the project. All ingest, validation and enrichment preservation actions are performed with the Autonomous Components described in section 4.2.

4.1 Repository

Each new batch of scanned newspaper pages must be ingested in our repository system, undergo a large number of quality checks and have access copies generated.

In keeping with the Minimal Effort Ingest model, we first ingest the batch of pages, and then perform the quality checks. Metadata is stored in DOMS, our Fedora Commons[5] 3.x based repository, whereas the data files are stored in our BitRepository[6]. We use Solr[7] to index the content of DOMS for our discovery platforms.

We store an additional object in DOMS, the batch object, which represents the batch of scanned pages, rather than any single page. In this object, we store PREMIS[8] Events detailing which actions and transformations have been performed on the batch. This information is also indexed by Solr.

4.2 Autonomous Components

We implemented the Autonomous Preservation Tool model in what we call Autonomous Components. Each component corresponds to a single action, such as “Ingest batch into repository” or “Schema-validate all XML files in batch”.

All autonomous components are Linux executables and have the following characteristics:

- Can query Solr for batch objects having specific combinations of PREMIS Events.
- Registers a component-specific PREMIS Event on the batch object after execution.

The current location of a batch in the workflow is determined by the set of PREMIS events present on the batch object - in other words which components have processed the batch so far. Each component knows which PREMIS events must be present or absent on a given batch for it to be ready to be processed by the component.

We have created *Tree Iterators* as a framework for autonomous components to handle batches in a storage-agnostic way. Tree iterators allow you to iterate through complex directory structures, whether in the repository or on disk, in a uniform way. With this framework, the autonomous components are able to work identically on batches not yet ingested, and batches inside the repository. This gives us great flexibility when testing, and allows us to easily rearrange which components should be run before ingest, and which should be run after.

5. CONCLUSIONS

We have shown that Autonomous Preservation Tools can be considered a viable alternative to a static workflow. We believe that Autonomous Preservation Tools should become a standard part of the digital preservationist’s toolbox, especially when using Minimal Effort Ingest.

6. FURTHER WORK

The State and University Library is currently in the process of replacing our Fedora based metadata repository. This will require a number of components to be reimplemented but we remain dedicated to the Minimal Effort Ingest and Autonomous Preservation Tools concepts.

During 2016 we will begin a project of receiving Danish born-digital newspapers. The principles described here will be carried further in this project.

7. ACKNOWLEDGEMENTS

We would like to thank our colleagues Kåre Fieldler Christiansen, Jette G. Junge and Gitte Bomholt Christensen for their help in developing both these concepts and this poster.

8. REFERENCES

[1] B. A. Jurik, A. A. Blekinge, and K. F. Christiansen. Minimal effort ingest. In Lee et al. [2].

[2] C. Lee, E. Zierau, K. Woods, H. Tibbo, M. Pennock, Y. Maeda, N. McGovern, L. Konstantelos, and J. Crabtree, editors. *Proceedings of the 12th International Conference on Digital Preservation.* School of Information and Library Science, University of North Carolina at Chapel Hill, 2015.

[3] Space Data and Information Transfer Systems. *ISO 14721:2012 Open Archival Information System (OAIS) - Reference Model.* The International Organization of Standardization, 2012.

[1] http://en.statsbiblioteket.dk/national-library-division/newspaper-digitisation/newspaper-digitization

[2] https://www.loc.gov/standards/mix/

[3] https://www.loc.gov/standards/mods/

[4] https://www.loc.gov/standards/alto/

[5] http://fedorarepository.org/

[6] http://bitrepository.org/

[7] http://lucene.apache.org/solr/

[8] http://www.loc.gov/standards/premis/

Born Digital 2016: Generating public interest in digital preservation

Sarah Slade
State Library Victoria
328 Swanston Street
Melbourne, VIC 3000, Australia
+61 (03) 8664 7383
sslade@slv.vic.gov.au

ABSTRACT

This poster describes the development and delivery of a week-long national media and communications campaign by the National and State Libraries of Australasia (NSLA) Digital Preservation Group to broaden public awareness of what digital preservation is and why it matters. Entitled *Born Digital 2016: collecting for the future*, this campaign will be linked with the 25th anniversary of the launch of the World Wide Web (6 August 2016) to gain maximum media exposure. The campaign focuses on the often unexpected benefits to the wider community of collecting and preserving digital material, rather than on the concept of loss which so often underpins arguments about why digital preservation is important.

Keywords

Digital Preservation, Communications campaign, NSLA Libraries, Media, Public engagement.

1. INTRODUCTION

The National and State Libraries of Australasia (NSLA) is the peak body for the ten National, State and Territory libraries of Australia and New Zealand. Each individual library is at a different stage in their digital collecting maturity. All are building and providing access to digital collections but only a few have active digital preservation systems and programs in place.

In July 2012, NSLA established a Digital Preservation Group (DPG) to identify best practice and collaborative options for the preservation of born digital and digitised materials [1].

When it was created, the DGP identified six priority work packages:

1. What is it and why? A statement on digital preservation and set of principles.
2. How well? A Digital Preservation Environment Maturity Matrix.
3. Who? A Digital Preservation Organisational Capability and Skills Maturity Matrix.
4. Nuts and Bolts: A common technical registry for NSLA libraries of file formats with software and hardware dependencies.
5. Collaboration and Partnership: A summary of opportunities for promotion and international representation and collaboration.
6. Confronting the Abyss: A business case for dedicated research into how to preserve difficult digital object types.

These work packages take into account the different stages of NSLA libraries in the adoption, development and implementation of digital preservation.

This poster focuses on work package 5 (Collaboration and Partnership). It describes the development and delivery of *Born Digital 2016: collecting for the future*, a five day national media and communications campaign across Australia and New Zealand to broaden public awareness of what digital preservation is and why it matters.

2. DEVELOPMENT

The message about digital collecting and preservation has generally focused on the amount of digital material being lost to future generations due to inadequate digital collecting practices and the lack of resources and systems. While all of this is important and true, the DPG felt that it was important to reframe the discussion with a more positive focus in order to achieve the aim of engaging the public and traditional media in this campaign.

As a result, *Born Digital 2016* will highlight the often unexpected benefits to the wider community of collecting and preserving digital material.

It was agreed that the most effective way to achieve this was with a collaborative, coordinated communications strategy across five themes—one for each day of the campaign. The daily theme provides an opportunity for national and local engagement with audiences through traditional and social media, and for individual libraries to hold events. The key messages for each theme will reinforce the role of NSLA libraries and all collecting institutions in digital collecting, preservation and access.

2.1 Themes

The five themes for the campaign were chosen to engage a broad range of community sectors and ages. Each theme provides a different focus for the public thinking about why digital material should be collected and preserved.

The five themes—Science and Space; Indigenous Voices; Truth and History; Digital Lifestyles; and Play—were chosen to encourage engagement, debate and media interest.

2.1.1 Science and Space

This theme highlights the importance of collecting and preserving scientific data to inform future thinking. A key example of this is the NASA space program's need to access and analyse data into the future, leading to the development of the foundations of digital preservation practice.

2.1.2 Indigenous Voices

This theme emphasises the vital role that indigenous media, archival collections and artefacts play in maintaining Indigenous culture and revitalising Indigenous language, particularly for communities with strong oral traditions.

2.1.3 Truth and History

This theme focuses on collecting and preserving online content and social media about political events in an objective manner, allowing communities to revisit and reshape notions of historical truth.

2.1.4 Digital Lifestyles

This theme looks at the storage of vast quantities of photographs, documents, memories and records on personal devices, home computers and in the cloud. It addresses how this vast volume of personal material is collated and kept safe and discoverable in the long term.

2.1.5 Play

This theme considers the role of digital games in the collective cultural memories of communities. It focuses on the complexity of preserving digital games—from old-school arcade machines to today's popular home-gaming platforms and the possibilities of immersive gaming in the future.

2.2 Timing

It is important that these types of campaigns are held at a time when the interest of the traditional media can be maximised. Examination of key national and state events, including election campaigns and major sporting events, identified dates to be avoided. The right time needed to both avoid these dates and link to a significant event relating to the digital world that the traditional media would already be interested in.

The week of 8–12 August 2016 met all of these requirements with Saturday 6 August 2016 marking the 25th anniversary of the launch of the World Wide Web and providing a valuable opportunity to leverage media promotion. The week also includes the International Day for Indigenous Peoples (9 August) and International Youth Day (12 August) which link directly with two of the week's themes.

2.3 Format and Experts

Key to the strategy is a high-profile expert speaker for each theme. These experts include scientists, journalists, academics and gaming and media personalities. They will be vodcast talking about their area of expertise and promoting discussion and debate about the importance of collecting and preserving digital material. The benefits that arise from this material being kept safe and made available will be a particular focus of these vodcasts.

Each NSLA library will deliver the vodcast for the theme of the day via their website. This will be accompanied by information about the theme and the library's digital activities and collections. To support this media partners will provide opportunities for radio and newspaper spots, including interviews with CEOs and digital preservation experts in each library. This will be complemented by a series of social media strategies for a digitally-driven campaign.

The overall communications and media strategy is coordinated by a multi-institutional Project Control Group, with expertise in digital preservation, communications/marketing, website and technology.

A working group in each NSLA library will also develop local events to be held at their institution focussing on at least one of the key themes.

3. DELIVERY AND IMPACT

Born Digital 2016 will run from 8–12 August 2016. The poster will include a summary of the activities undertaken and highlight the level of media and public engagement achieved.

4. ACKNOWLEDGMENTS

The Author would like to thank Serena Coates, Steve Knight, Nerida Webster, Stephen Sayers, Golda Mitchell, Lynn Benson, Aimee Said and Kate Irvine for their work on the development of *Born Digital 2016*. Thanks also go to NSLA and the individual NSLA libraries for their involvement in this campaign.

5. REFERENCES

[1] NSLA Digital Preservation Group. Available at http://www.nsla.org.au/projects/digital-preservation. Accessed 25 April 2016

Building Archival Storage Component Using New Open Standard Infrastructure

Dariusz Paradowski
National Library of Poland
al. Niepodleglosci 213
02-086 Warsaw Poland
+48 22 608 26 17
d.paradowski@bn.org.pl

ABSTRACT

The National Library of Poland (NLP) makes use of emerging open standard of digital magnetic tape structure - LTFS to build simple, efficient, economic, scalable and safe archival storage component for institutional repository.

Keywords

Digital archival storage; open standard; long time digital preservation; digital library

1. INTRODUCTION

For over ten years the National Library operates comprehensive system for supporting digitization of library materials, processing created digital files, presentation and storage. The system called Repository has a number of functional modules responsible for multitrack workflow, import from various sources (scanning, digital legal deposit), conversion of data formats and metadata, generating derivatives (jpg, OCR, etc.) presentation and access management, storage and finally archiving. The current generation of the repository immediately after the introduction of the source files to the system places them in an object archive. It is a commercial appliance in the form of licensed software running on a multi-node cluster of dedicated servers connected to the disc array.

In the beginning it was a very useful solution:

- it provided a guarantee that extremely important archiving module will be reliable - it was a commercial solution verified in considerable applications,
- cluster equipped with load balancing efficiently acquired and provided files,
- it supplied manageable WORM functions,
- by compression it optimally used disc array capacity, ensuring that the current amount satisfied the needs for a longer period of time.

2. ISSUE

After a few years the rate of digitization in BN, was greatly raised which, accompanied by quality improvement, resulted in a rapid increase of the bytestream of files entering the repository.

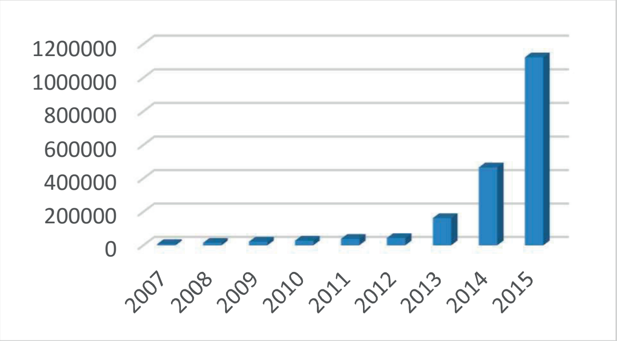


Figure 1. Items available in polona.pl

As a result, there appeared drawbacks of the used archive:

- insufficient performance - expected increase would require changes to the architecture of application transmitting the data to the archive and also hardware upgrade of the appliance (dedicated servers and arrays),
- rapid depletion of capacity - the further operation would require costly expansion of the storage array and also expensive purchase of licenses for the appliance archive
- increasing energy consumption of the solution.

3. CHALLENGE

To solve these problems reasonably, different archive module was needed. Essential requirements of equal importance were defined:

- low cost storage
- high, easily scalable performance
- high, easily scalable capacity
- safety

These were then turned into more practical ones:

- low cost storage
    - low cost of capacity per byte
    - no expensive license capacity
    - low energy consumption
- high, easily scalable performance
    - horizontal scaling possible without rebuilding the infrastructure (just extension) and software
    - vertical scaling possible without software change
- high, easily scalable capacity
    - horizontal scaling possible without rebuilding the infrastructure (just extension) and software
    - vertical scaling possible without software change
- safety
    - open standards of storage writing
        - the system can not depend on single manufacturer
        - data must be readable outside environment of the archive
        - metadata must be human-readable
    - recognized standards
    - damage to any part of the data can not prevent the reading of data undamaged

4. FACTORS CONSIDERED

4.1 Choice of Carrier

Aside from stone and paper the magnetic tape is best recognized carrier that has very long shelf life. Tape has exceptionally low cost per unit of capacity. It is essential to use an open and yet recognized standard. Linear Tape Open (LTO) is an open standard supported by many major manufacturers, it also has defined roadmap for development. The new LTO7 standard appeared on the market at the end of 2015 and has a capacity of 6TB. This is enough to avoid necessity to purchase another expansion frame for the automatic tape library currently used in NLP in the foreseeable future. It has long time to the end of support and moreover will be readable by next two generations of drives [1].

Table 1. RAID10 vs independent carriers vulnerability

	Data loss %		
No of carriers destroyed	hard disc drives, RAID 10, (4 mirrored = 8 drives) best case	2 sets of 4 independent tapes worst case	2 sets of 4 independent tapes best case
1	0	0	0
2	0	25	0
3	0	25	0
4	0	50	0
5	100	50	25
6	100	75	50
7	100	75	75
8	100	100	100

Comparison of tape to disc storage. Table 1 shows comparison of the sensitivity of the system disk in a very robust and expensive version of RAID 10 compared with a set of independent carriers comprising two copies of the data. With minor injuries RAID gives greater protection than the worst case for independent media and the same as the best case. It is worth noting that the worst case is relatively unlikely (damage to same data on different tapes) and the best case gives better results than RAID. In particular, the destruction of more than 50% of the media RAID causes a loss of 100% of the data, while the media independent only 25-50%. It is also worth noting that the price of a unit capacity of good quality media is much lower for tape cartridges. It is also important from the point of long time preservation that in case of danger cartridges may be easy removed from tape library and transported which is much more complicated for hard drives.

Sequential recording on tape cartridges, which corresponds to scenarios of archival usage is performed with a very high speed. The performance of the system can be easily multiplied by increasing the number of drives and the capacity by increasing number of tapes.

4.2 Choice of Filesystem

Linear Tape File System (LTFS) meets the requirement of the system to be open, it is supported by several leading manufacturers, developed in the mature form and present on the market for several years. This year – 2016 LTFS became adopted as standard ISO / IEC [2]. Record in LTFS can be read on another device from another manufacturer, without the need to reconstruct an environment where it was saved. Moreover, basic software solutions - allowing the use LTFS on a single drive are available as open source by many hardware manufactures.

4.3 Developed Methods

The system is to serve as a disaster recovery solution, so the assumption that there will survive a random subset of cartridges implicates the requirement that each object will be stored on one medium. A complete object is understood as: a unique object identifier, all the metadata and structure of the object in the form of (human readable) XML METS and all source content files of the object.

A carrier that most of the time is kept offline allows to postulate an idea of avoiding backwards error propagation. Once saved, the object in the archive is never to be changed. Any change will be treated as formation of a new version of the whole object. Information about the location of the next version (barcode of the cartridge) will be placed in the database system. It freed us from designing a complex and unreliable predictions of reserved free space on the tape needed to create new versions of files (that way would be also very inefficient considering linear nature of the tape recording).

The challenge was the metadata, which, in the national library reality are subject to frequent revisions. This problem was solved by independent archiving the updated metadata of all objects through saving the entire database of the system in an XML file (with checksums). Likewise objects each database copy is to be kept without adjustments and with versioning instead (version and time markers apply).

Thus, if after the disaster, a random set of tape cassettes has been discovered, it is enough to find the latest version of the database which allows to quickly find the latest versions of objects.

To economically save objects of random, often large size on the discrete space on the tapes, it is necessary to optimize it. The archive will temporarily gather on it's own disc buffer, objects supplied by the repository that are ready for archiving. Choosing from this pool archive will construct packages of these objects of a size as close as possible to the capacity of a single tape cartridge to write it at once.

Contradictory parameters such as the maximum allowable time of the object in the buffer and the minimal tape capacity loss will be fine-tuned on the basis of statistics. After achieving at optimal thresholds, the tape will be written only once and never changed. This approach ensures maximization of write speed and durability of the tape. This also allows to overcome the incompatibility between WORM and LTFS by switching write protection tab on the tape cartridge further improving safety of the archive.

5. REFERENCES

[1] Buffington, J. 2016. *Analyzing the Economic Value of LTO Tape for Long-term Data Retention.* http://www.lto.org/wp-content/uploads/2014/06/ESG-WP-LTO-EVV-Feb_2016.pdf

[2] *ISO/IEC 20919:2016 Information technology — Linear Tape File System (LTFS) Format Specification* (First edition) https://www.iso.org/obp/ui/#iso:std:iso-iec:20919:ed-1:v1:en

Caring For Live Art That Eludes Digital Preservation

Louise Curham
Centre for Creative & Cultural Research, Faculty of Arts & Design
University of Canberra, ACT 2601, Australia
+61 412 271 612
louise.curham@canberra.edu.au

ABSTRACT

This poster outlines my research on strategies of re-enactment to keep alive artworks that rely on performance. While digital documentation for some of these works circulates, the live nature of the works means they evade meaningful digitisation. In an artist/archivist collaboration, Teaching and Learning Cinema, myself and colleague Dr Lucas Ihlein have evolved three principal ways to bring these works from the original artists through to future generations – direct engagement with the original artist, extensive documentation of the re-enactment process and the formulation of new 'expressive' instructions.

This approach resonates with a newly ignited discussion in Australia about how the conservation profession can effectively reach beyond institutions to communities. This work suggests that empowering communities to find their own solutions to intergenerational transmission means the process of preservation becomes part of the cultural product, a preservation of doing.

Keywords

Archives, cultural collections, intangible cultural heritage, re-enactment, contemporary art.

1. INTRODUCTION

In the past decade, the Australian artist/archivist collaboration Teaching and Learning Cinema (TLC) has been working with live artworks made by artists in previous generations. These are artworks that were made to be experienced live and made without concern about how future generations might be able to experience them. While institutions increasingly collect and preserve works that include ephemeral elements like video or degrading materials, the situation is different for works that are made to be experienced live. In 2016, the conservation profession in particular is engaging with this problem. Discussions about the preservation of embodied knowledge that takes the form performance and other kinds of ephemeral art are the focus of two international conservation meetings this year, IIC in Los Angeles and a conservation profession symposium in German [1]. This communicates the urgency and interest about the problems of keeping alive cultural heritage whose essence is other than a tangible object.

This word 'essence' links us to digital preservation where we expect an essence attached to performance layers and carriers. In the case of the live art TLC is concerned with, the works are scattered both physically and intellectually. TLC's experience points to a way to bring this scattered essence together. While TLC's work engages with resolutely analogue examples, it sheds light on how such a process could also occur with digital essences that may defy what have become expected digital preservation pathways.

2. EXPANDED CINEMA AND THE PRESERVATION PROBLEM

TLC's work is concerned with a subset of live art, film performance artworks known as Expanded Cinema. These works combined experimental film with live performance. Their lineage in 20th century art lies in performance art, conceptual art and early media art and installation [2].

While part of these Expanded Cinema works consist of tangible objects such as 16mm film or super 8, there are no instructions for the work and the knowledge about it is distributed, for example between the original artist, film archives and other collections.

To illustrate this, in 2013, TLC visited British film artist Malcolm Le Grice who had decided it was time to 'train a stand-in' [3] for his work *Horror Film 1* (1971), a work for multiple 16mm projections and performer.

At first glance, we could presume that *Horror Film 1* is safe for the future – Le Grice still performs it, social media captures his recent past performances, a film archive has video documentation of it along with the film raw materials and other archives hold programs, photographs and correspondence about its early performances and material about the scene in London it emerged from. Yet none of these material traces about *Horror Film 1* can stand in for the experience of the work itself.

3. TEACHING AND LEARNING CINEMA RE-ENACTMENTS

TLC began re-enacting Expanded Cinema works so that we could experience them for ourselves. The works that we are drawn to have emerged from the scene around the London Film Makers' Co-op in the late 1960s and early 1970s. While there has been an international resurgence of interest in these works[1] and the original artists continue to perform them, there is little access to performances of them for Australian audiences.[2] Our distance from London contributed to the logic of re-enacting the works in the first place [4], re-enactment making little sense if we had ready access to performances of the works by the original artists.

From our evolving process, three consistent approaches have emerged: direct engagement with the original artist, extensive documentation of the process and formulating 'expressive' instructions.

Our 2009 project on British artist Guy Sherwin's *Man With Mirror* (1976) sets out this process of direct engagement with the original artist – a straight forward process of gauging his interest and forging connections with him. His positive response led to him stepping us through the work during a visit he made to Australia in 2008. This direct transmission from Sherwin to TLC made it possible for us to make sense of the resources brought together from our research eg we found diagrams and other descriptions of the series of movements the work requires that are performed with a mirror. It was not until we spent the short time under Guy's tuition that we could make real sense of this material.

The second part of our approach involves extensive documentation of our process using a blog to record diary-type entries of our experience and to capture knowledge of the structure and technical details of the work as they emerge. Examples of entries include drawings, photographs and digitised archival material we locate in our research along with reflections on the work as it unfolds. This has several impacts. It captures our decision points, where inevitable deviations from the original work occur. These points become critical for us as part of the new artwork we create through the re-enactment, making transparent where and why these decision points have occurred. An example of this is the decision to include two performers in our re-enactment of *Man With Mirror* – TLC's re-enactment became *(Wo)Man With Mirror*. This apparently minor change shifts the emphasis substantially from Sherwin's original – for example audiences read the piece as a commentary on male female relations, not relevant in Sherwin's original. In capturing our decision points, there is a record of how our knowledge about the work unfolded, akin to the reversible treatments in preservation.

For *(Wo)Man with Mirror*, we then captured this knowledge in the form of a user's manual that set out context for the work, background about Sherwin along with step-by-step instructions to put the work together. In 2016, we worked with a young artist, Laura Hindmarsh, to use the user's manual. This highlighted its gaps as 'expressive instructions', to use American philosopher Richard Sennett's phrase, points where the manual failed to overcome the gap between instructive language and the body [5].

4. TLC'S RE-ENACTMENTS AS PRESERVATION AND AS A PLATFORM FOR BUILDING COMMUNITY

TLC's approach resonates with a newly ignited discussion in Australia about how preservation services can effectively reach communities beyond institutions. In 1995, Australia was ground-breaking in embracing a national preservation policy. A recent call to revisit this policy in part responds to the situation where preservation work occurs predominantly within institutions and proposes measures to expand this work into the wider community [6]. The proactive labour of re-enactment puts the available resources to work to make an iteration of these artworks, behaving as a practical form of preservation.

The *(Wo)Man With Mirror* user's manual engaged another artist, expanding the community that cares about this work and now engages in seeing it survive in the future. The user's manual points the way to the process of re-enacting the work as one of community building. This suggests that part of the solution to the problem of preservation is for communities to care for their important stuff themselves. The work of TLC is one example of how we might transmit our work from one generation to the next in an iterative process where the work is an opportunity for community building in and of itself. The work is no longer the invisible professional work of the conservator but an active engagement with the work and the documentation of that engagement becomes both the work and its preservation – a preservation of doing.

5. ACKNOWLEDGMENTS

The work of the Teaching and Learning Cinema is a collaboration between Dr Lucas Ihlein, lecturer Media Arts, University of Wollongong and the author. The author wishes to acknowledge the instrumental role of artists Malcolm Le Grice and Guy Sherwin in the re-enactment work of Teaching and Learning Cinema and the recent contribution of artist Laura Hindmarsh.

6. REFERENCES

[1] Verband der Restoraten, 'Collecting and Conserving Performance Art', international symposium, 9-11 June 2016. Accessed 20 April 2016 at http://www.restauratoren.de/termine-details/2021-collecting-and-conserving-performance-art.html and IIC 2016, 'Saving the Now: crossing boundaries to conserve contemporary works', congress of the International Institute for Conservation of Historic and Artistic Works. Accessed 20 April 2016 at https://www.iiconservation.org/node/5586.

[2] Ihlein, L. (2005). Pre-digital New Media Art. *Realtime* 66. Accessed 20 April 2016 at http://www.realtimearts.net/article/issue66/7779.

[3] Le Grice, M. (2001). Improvising time and image. *Filmwaves, 14 (Nov)*, p. 17.

[4] Ball, S. (2016). Beyond the cringe: Australia, Britain and the post-colonial film avant-garde. *Senses of Cinema, 78 (May)*. Accessed 20 April 2016 at http://sensesofcinema.com/2016/british-experimental/post-colonial-film-avant-garde/.

[5] Sennett, R. (2008). *The Craftsman*. London: Penguin Books, p. 179.

[6] Sloggett, R. (2016). A national conservation policy for a new millennium—building opportunity, extending capacity and securing integration in cultural materials conservation. *Bulletin of the Australian Institute for the Conservation of Cultural Materials, 36(2), 79-87.*

[1] In 2002, a major retrospective film program and research project entitled *Shoot Shoot Shoot, The First Decade of the London Film-Makers' Co-operative and British Avant-Garde Film 1966-76*, launched at Tate Modern and embarked on a world wide tour.

[2] An exception to this is work by Australian artist group OtherFilm who toured Guy Sherwin, Malcolm Le Grice and other moving image artists to Australia from 2008-10.

Consortial Certification Processes – The Goportis Digital Archive. A Case Study

Yvonne Friese
ZBW – Leibniz Information Centre
for Economics
Düsternbrooker Weg 120
D-24214 Kiel
+49 431 8814610
y.friese@zbw.eu

Thomas Gerdes
ZBW – Leibniz Information Centre
for Economics
Neuer Jungfernstieg 21
D-20354 Hamburg
+49 40 42834311
t.gerdes@zbw.eu

Franziska Schwab
TIB – German National Library of Science and Technology
Welfengarten 1B
D-30167 Hannover
+49 511 76219073
franziska.schwab@tib.eu

Thomas Bähr
TIB – German National Library of Science and Technology
Welfengarten 1B
D-30167 Hannover
+49 511 76217281
thomas.baehr@tib.eu

ABSTRACT

The Goportis Consortium consists of the three German National Subject Libraries. One key area of collaboration is digital preservation. The Goportis Digital Archive is jointly used by the consortial partners. As part of their quality management the partners strive to obtain certifications for trustworthy digital repositories. The Goportis Consortium successfully applied for the Data Seal of Approval (DSA) [1] and is currently working on the application for the nestor Seal [2].

The poster illustrates the collaboration of the Goportis partners during the certification process (distribution of tasks, time frame, etc.). This way it could serve as best-practice example for other institutions interested in consortial certification.

Keywords

Certification; Consortium; Consortial Certification; Audit; Digital Archive; Digital Preservation

1. INTRODUCTION

In Germany there are three National Subject Libraries working together in the Goportis Consortium: the German National Library of Science and Technology (TIB, Hannover), the German National Library of Medicine – Leibniz Information Centre for Life Sciences (ZB MED, Cologne/Bonn) and the German National Library of Economics – Leibniz Information Centre for Economics (ZBW, Kiel/Hamburg). To ensure the preservation of their digital contents, in 2010 they jointly founded the Goportis Digital Archive. The archive is based on Rosetta, the Ex Libris software. TIB is the main licensee, ZBW and ZB MED have sublicenses. The computing centre for the Digital Archive is hosted by TIB. That is why TIB takes care of system administration and the general settings.

For the Goportis partners the certification of their digital archives is part of their quality management, since all workflows are evaluated. Beyond that, a certification seal signals to external parties, like stakeholders and customers, that the long-term availability of the data is ensured and the digital archive is trustworthy. There is an array of different options and programs for certification. Goportis chose to follow the path of the European Framework for Audit and Certification of Digital Repositories [3]. In 2015, TIB and ZBW successfully applied for the Data Seal of Approval as basic first-level certification. At present, we are working on the application for the nestor Seal, an extended second-level certification based on DIN 31644. ZB MED is currently not involved in the certification process.

The poster focuses on the certification process for the DSA and the nestor Seal. It illustrates the workflows and distribution of tasks within the Goportis Consortium. Additionally, it offers a time frame for the certification process. There is also an outlook on future plans for third-level certification. The information provided could be reused by other consortia who want to jointly certify their digital archives.

2. DATA SEAL OF APPROVAL

The Data Seal of Approval was established in 2008 and is based on 16 guidelines (i.e. criteria). It is an externally reviewed self-audit which is publically available [4]. The DSA aims at ensuring "that archived data can still be found, understood and used in the future" [5]. For each applicable guideline there are four levels of compliance:

- No: We have not considered this yet.
- Theoretical: We have a theoretical concept.
- In progress: We are in the implementation phase.
- Implemented: This guideline has been fully implemented for the needs of our repository.

2.1 Distribution of Tasks

In general, all three German National Subject Libraries are equal partners within the Goportis Consortium. For digital preservation, though, TIB is the consortium leader, since it is the software licensee and hosts the computing centre. Due to the terms of the DSA—as well as the ones of the nestor Seal—a consortium cannot be certified as a whole, but only each partner individually. This constellation determined how the consortium applied for the DSA: In principle, each partner drew up its own application. However, for some aspects of the certification ZBW had to refer to the answers of TIB, which functions as its service provider.

Beside these external requirements the Goportis partners organized the distribution of tasks on the basis of internal goals as well. They interpreted the certification process as an opportunity to get a deeper insight in the workflows, policies and dependencies of the partner institutions. That is why they analyzed the DSA guidelines together and established a time frame. Moreover, they discussed the progress of the application process regularly in telephone conferences and matched their answers to each guideline. As a positive side effect, this way of proceeding strengthened not only the ability of teamwork within the consortium. It also led to a better understanding of the guidelines and more elaborate answers for the DSA application.

The documentations for the DSA were created in more detail than recommended in order to facilitate further use of the documents for the nestor Seal.

Section 2.2 gives some examples of DSA guidelines for which ZBW depended upon the consortium leader TIB. It explains as well which guidelines differed for each institution.

2.2 DSA Guideline Examples

2.2.1 Guidelines depending on the consortium leader TIB

Guideline 6: "The data repository applies documented processes and procedures for managing data storage." Dependence: The permanent storage of the Goportis Digital Archive is administrated by TIB in Hannover.

Guideline 8: "Archiving takes place according to explicit workflows across the data life cycle." Dependence: The general information about the workflows within Rosetta was described by the software licensee TIB. Notwithstanding this, ZBW provided additionally a detailed answer for its own collections.

Guideline 11: "The data repository ensures the integrity of the digital objects and the metadata." and guideline 12: "The data repository ensures the authenticity of the digital objects and the metadata." Dependence: Workflows that ensure integrity and authenticity depend on the Rosetta system. ZBW completed its own description with a reference to the information provided by TIB.

2.2.2 Guidelines that differ for each institution

Several guidelines had to be answered by each Goportis partner individually, for example, criteria concerning the archived digital data, data users, data formats and metadata (guidelines 1-3, 10). The same is true for criteria regarding the institution, its mission, the responsibility for the digital data as well as the OAIS compliance of the whole digital archiving workflow (guidelines 4, 5, 9, 13-16).

2.3 Time Frame

The certification process extended over six months. In each Goportis institution one employee was in charge of the DSA certification process. Other staff members provided additional special information about their respective areas of work. This included technical development, data specialists, legal professionals, team leaders and system administration (TIB only).

Table 1: Person months for the certification process

Institution	Person Responsible	Other Staff
TIB	~ 3	~ 0.25
ZBW	~ 1.5	~ 0.1

The time estimate described in Table 1 adds up to approx. 3.25 (TIB) resp. 1.6 (ZBW) person months. Included are twelve telephone conferences (90 minutes each), which were held in order to coordinate the consortial certification process.

3. OUTLOOK: NESTOR SEAL & THE THIRD LEVEL

The nestor Seal represents the second level of the European Framework for Audit and Certification of Digital Repositories. With its 34 criteria it is more complex than the DSA [6]. It also requires more detailed information, which makes it necessary to involve more staff from different departments. Based on the good experiences of the DSA certification, the Goportis partners TIB and ZBW plan to acquire the nestor Seal following the same way of proceeding. The DSA application has prepared the ground well for this task, since important documents, like policies, have already been drafted. That is why we estimate the application process for the nestor Seal to require approx. 12 (TIB) resp. 4 (ZBW) person months.

Having completed the application for the nestor Seal, the Goportis Consortium will discuss the application for a third-level certification based on ISO 16363, which requires an external audit. A consortial certification at this level will certainly pose special challenges, e.g. because some aspects of the digital preservation are provided for all consortial partners by TIB, whereas workflows and preservation decisions are the individual responsibility of each Goportis partner.

4. REFERENCES

[1] *Data Seal of Approval*. URL= http://datasealofapproval.org/en/information/about/.

[2] *nestor Seal*. URL= http://www.langzeitarchivierung.de/Subsites/nestor/DE/nestor-Siegel/siegel_node.html.

[3] *European Framework for Audit and Certification of Digital Repositories*. URL= http://www.trusteddigitalrepository.eu/Trusted%20Digital%20Repository.html.

[4] *Assessment Data Seal of Approval TIB*. URL= https://assessment.datasealofapproval.org/assessment_157/seal/pdf/.
Assessment Data Seal of Approval ZBW. URL= https://assessment.datasealofapproval.org/assessment_158/seal/pdf/.

[5] Cf. reference [1].

[6] Schwab, F. 2016. TIB prepares for certification of its digital preservation archive according to DIN 31644 (nestor Seal). In *TIB-Blog* (24/6/2016). URL= https://tib.eu/start-certification-nestor-seal.

Digital Preservation with the Islandora Framework at Qatar National Library

Armin Straube
Qatar National Library
PO Box: 5825
Doha – Qatar
Tel. (+974) 4454 1551
astraube@qf.org.qa

Arif Shaon
Qatar National Library
PO Box: 5825
Doha – Qatar
Tel. (+974) 4454 1523
ashaon@qf.org.qa

Mohammed Abo Ouda
Qatar National Library
PO Box: 5825
Doha – Qatar
Tel. (+974) 4454 7190
maboouda@qf.org.qa

ABSTRACT

This poster outlines how Qatar National Library builds a versatile multi-purpose repository that will provide digital preservation solutions to a wide range of national stakeholders and use cases.

Keywords

Digital preservation; Islandora; Institutional repository

1. BACKGROUND

The Qatar National Library (QNL)[1] project was established in November 2012 as a non-profit organization under the umbrella of Qatar Foundation for Education, Science, and Community Development (QF)[2]. The library supports Qatar on its journey from a carbon-based to a knowledge-based economy by providing resources to students, researchers and the community in Qatar. The wider mission of QNL is 'to spread knowledge, nurture imagination, cultivate creativity, and preserve the nation's heritage for the future.'

QNL collects and provides access to global knowledge relevant to Qatar and the region. It also collects, preserves and provides access to heritage content and materials. From an operational standpoint QNL has three functions: a national library for Qatar, a university and research library to support education and research at all levels, and a public library to serve the metropolitan area.

All these functions are increasingly fulfilled in a digital way. Addressing issues of digital preservation has therefore become a cornerstone of QNLs operational remit. As a national library, QNL also recognizes an obligation to support other Qatari institutions from the cultural, research and scientific domains as well as other internal and/or external enterprise systems. While many of these institutions host a wealth of digitized and born-digital content, including a variety of research data or output, their preservation over the longer term has so far not been properly addressed. QNL aims to develop digital preservation solutions for both its own needs and for those of partner institutions.

2. QNL DIGITAL PRESERVATION STRATEGY

The QNL digital preservation strategy has been formulated to build a trustworthy digital repository on the basis of established standards in digital preservation and includes the certification of its achievements. The strategy supports a wide range of existing digital collections, including digitised cultural heritage collection[3], research data/output generated by Qatari academic/research institutions, and audio/visual materials hosted locally at QNL, as well as various other collections to be established in future.

In general, the QNL preservation strategy is underpinned by a number of guiding principles that serve as benchmarks for the library's development of its digital preservation efforts and inform its decision making process:

- **Accessibility** - permanent accessibility and usability of all preserved digital content is the principal goal of the QNL digital repository.
- **Integrity** - ensuring the bitstream preservation of archived material is a basic requirement. QNL will take appropriate measures like the regular verification of checksums, multiple storage redundancies and the monitoring and exchanging of storage hardware.
- **Persistent identifiers** - all digital objects will be referenced by a (globally) unique and persistent identifier.
- **Metadata** - QNL will capture technical metadata about all digital objects ingested for preservation and will record information about preservation actions and events using PREMIS.
- **Preservation planning and risk assessment** - all objects ingested in the QNL digital repository will undergo a risk assessment, the result of which will form the basis for decision making on preservation action. The assessment is to be updated and checked regularly to account for technological changes and related economic factors.
- **Standards compliance and trustworthiness** – the QNL digital repository is to be built on the basis of established standards in digital preservation (ISO 14721, ISO 16363, DIN 31644) to ensure longevity and trustworthiness.
- **Development and research via collaboration**- QNL recognises that the complexity and diversity of challenges associated with long term digital preservation is beyond the scope of any single organization. The library will therefore monitor the state of the art in long-term digital preservation and seek to participate in collaboration and research at both national and international levels to facilitate future development of the digital preservation infrastructure as applicable.

Due to the changing nature in the area of digital preservation and the rapid development of services and content acquisition at QNL, the preservation strategy will be reviewed and revised in 2018 at the latest.

[1] http://www.qnl.qa/
[2] https://www.qf.org.qa/
[3] http://www.qnl.qa/collections/aihl

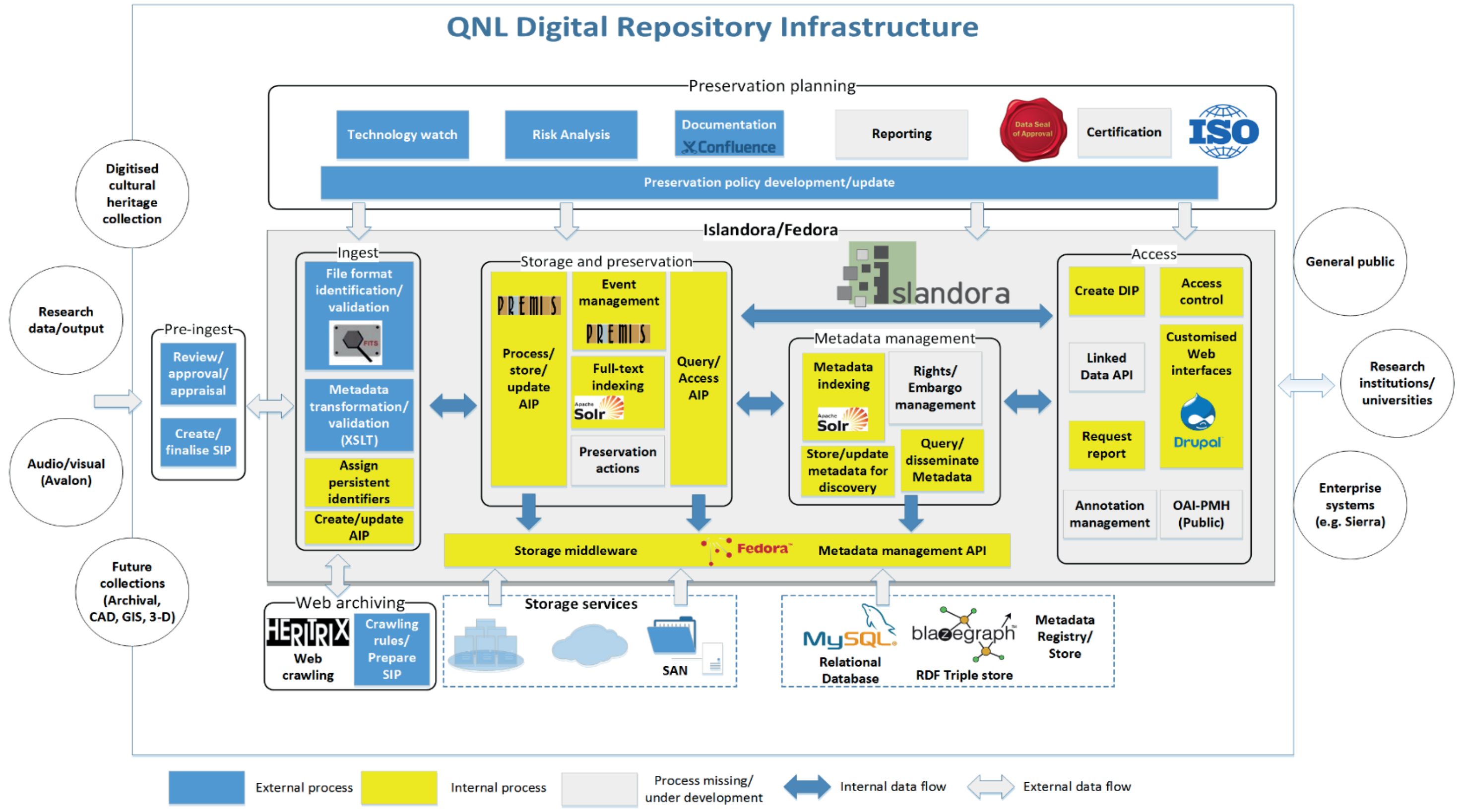


Figure 1: QNL Digital Repository Infrastructure

3. QNL DIGITAL REPOSITORY IMPLEMENTATION

The digital repository of QNL was set up in January 2016 to implement the QNL digital preservation strategy. As illustrated in Figure 1, the overall infrastructure for the QNL digital repository is based on Islandora[4] (version 7.x-1.6), an open-source digital content management system, integrated with Fedora Commons[5] (version 3.8.1) as the underlying repository technology. The architecture of Islandora is based on the Drupal[6] framework that allows different functions of a repository to be developed as Drupal modules, commonly referred to as the Islandora solution packs. The advantages of this modular architecture include customization, further development and integration with third party software.

Islandora provides a range of Drupal modules that support some of the important preservation functions, such as the management of persistent identifiers, the support of PREMIS or the integration of file format identification tools, such as FITS. In addition, the capacity of the Islandora modules can be enhanced by integrating with external preservation solutions – e.g. the Archidora[7] module that integrates Archivematica[8] with Islandora. The evaluation/implementation/deployment of these modules is currently under way. In general, Islandora access modules, underpinned by a uniform preservation framework, can be customized to serve a wide range of use cases and be adapted to the need of institutions other than QNL.

While QNL will dedicate its development resources to Islandora, solutions outside this framework can and will also be utilized where applicable. For example, the audiovisual collection of QNL is made available via the Avalon platform, representing the first deployment of this system in the Middle East.

At present, QNL digital repository mainly stores image based objects (digitized books, maps, photos etc.) in both tiff and jpeg 2000 formats, audio-visual collections in mp4 and wav, and web archives in warc format via integration with Heritrix[9], an external web crawling tool. Beyond these objects types, the repository is capable of providing bit stream preservation for any digital object and is under development to provide additional support for a wider range of digital objects and different metadata standards.

In addition, descriptive metadata are both stored in the repository and in the library's Sierra catalogue. Technical systems and workflows are documented in a wiki. QNL has its own dedicated storage infrastructure with tiered storage (hard drives and tape library) and a regular data backup schedule. A policy driven data management is used and multiple redundancies are kept.

4. FUTURE WORK

Future digital collections to be ingested into the QNL digital repository include archival material, GIS and CAD files, 3D scans of museum objects, and databases. The repository will be developed to be scalable to handle increasing volumes of content.

In addition, the library will develop a file format policy, formalizing its current implicit practice, which will enhance the basis of its risk assessment.

QNL aims for certification as a trusted digital repository and will apply for the Data Seal of Approval[10] in 2018 at the latest.

[4] http://islandora.ca/
[5] https://wiki.duraspace.org/display/FEDORA38/Fedora+Repository+3.8.1+Release+Notes
[6] https://www.drupal.org/
[7] https://wiki.duraspace.org/display/ISLANDORA715/Archidora
[8] https://www.archivematica.org/en/
[9] https://webarchive.jira.com/wiki/display/Heritrix
[10] http://www.datasealofapproval.org/en/

Establishing a generic Research Data Repository: The RADAR Service

Angelina Kraft
Technische Informationsbibliothek (TIB) German National Library of Science and Technology
Welfengarten 1 B
D-30167 Hannover, Germany
+49 (0)511 762 14238
angelina.kraft@tib.eu

Matthias Razum
FIZ Karlsruhe – Leibniz Institute for Information Infrastructure
Hermann-von-Helmholtz-Platz 1
D-76344 Eggenstein-Leopoldshafen, Germany
+49 (0)7247 808 457
matthias.razum@fiz-karlsruhe.de

Jan Potthoff
Karlsruhe Institute of Technology (KIT)
Hermann-von-Helmholtz-Platz 1
D-76344 Eggenstein-Leopoldshafen, Germany
+49 (0)721 608 25 666
jan.potthoff@kit.edu

ABSTRACT

Science and its data management are in transition. And while the research data environment has become heterogeneous and the data dynamic, funding agencies and policy makers push towards findable, accessible, interoperable and reuseable (= FAIR) research data [1]. A popular issue of the management of data originating from (collaborating) research infrastructures is their dynamic nature in terms of growth, access rights and quality. On a global scale, systems for access and preservation are in place for the big data domains (e.g. environmental sciences, space, climate). However, the stewardship for disciplines of the so-called long tail of science remains uncertain. This poster gives the impression of an interdisciplinary infrastructure facilitating **research data archival and publication.**

The **RADAR - Research Data Repository -** project strives to make a decisive contribution in the field of long tail research data: On one hand it enables clients to **upload, edit, structure and describe (collaborative) data in an organizational workspace**. In such a workspace, administrators and curators can manage access and editorial rights before the data enters the preservation and optional publication level. Data consumers on the other hand may **search, access, download and get usage statistics** on the data via the RADAR portal. For data consumers, findability of research data is of utmost importance. Therefore the metadata of published datasets can be harvested via a local **RADAR API** or the DataCite Metadata Store.

Being the proverbial **“transmission belt” between data producers and data consumers**, RADAR specifically targets researchers, scientific institutions, libraries and publishers. In the data lifecycle, RADAR services are placed in the “Persistent Domain” of the conceptual data management model described in the “domains of responsibility”[2]. These domains of responsibility are used to show duties and responsibilities of the actors involved in research data management. Simultaneously, the domains outline the contexts of shared knowledge about data and metadata information, with the goal of a broad reuse of preserved and published research data.

RADAR applies different preservation and access strategies for open vs. closed data:

- For open datasets, RADAR provides a Digital Object Identifier (DOI) to enable researchers to clearly reference data. The service offers the publication service of research data together with format-independent data preservation for at least 25 years. Each published dataset can be enriched with discipline-specific metadata and an optional embargo period can be specified.
- For closed datasets, RADAR uses handles as identifiers and offers format-independent data preservation between 5 and 15 years, which can also be prolonged. By default, preserved data are only available to the respective data curators, which may selectively grant other researches access to preserved data.

With these two services, RADAR aims to meet demands from a broad range of research disciplines: To provide a secure, citable data storage and citability for researchers which need to retain restricted access to data on one hand, and an e-infrastructure which allows for research data to be stored, found, managed, annotated, cited, curated and published in a digital platform available 24/7 on the other.

E-research projects often require comprehensive collaborative features. These include data storage, access rights management and version control. RADAR possesses a **modular software architecture** based on the e-research infrastructure eSciDoc Next Generation. The data storage is managed by a repository software consisting of two parts: A back end addresses general tasks such as storage access, **bitstream preservation** and regular reports on data integrity, whereas the front end implements RADAR-specific workflows. Front end workflows include various data services: Metadata management, access control, data ingest processes, as well as the licensing for re-use and publishing of research data with DOI. Archival Information Packages (AIP) and Dissemination Information Packages (DIP) are provided in a **BagIt-structure**[3] in ZIP container format. As part of the import/export strategy, an API for RADAR will be provided. The API allows the import/export of data as well as metadata.

The **RADAR API enables users to integrate the archival backend into their own systems and processes**. Another option is to install the RADAR software locally. The customer may choose to only deploy the management and User Interface layer, while archiving the data in the hosted RADAR service via the API, or to run everything locally. Additionally, there is the option to run the complete software stack locally and use the hosted RADAR service as a replica storage solution.

RADAR is developed as a cooperative project of five research institutes from the fields of natural and information sciences. The technical infrastructure for RADAR is provided by the FIZ Karlsruhe – Leibniz Institute for Information Infrastructure and the Steinbuch Centre for Computing (SCC), Karlsruhe Institute of Technology (KIT). The sustainable management and publication of research data with DOI-assignment is provided by the German National Library of Science and Technology (TIB). The Ludwig-Maximilians-Universität Munich (LMU), Faculty for Chemistry and Pharmacy, and the Leibniz Institute of Plant Biochemistry (IPB) provide the scientific knowledge and specifications and ensure that RADAR services can be implemented to become part of the scientific workflow of academic institutions and universities.

KEYWORDS

Research Data Infrastructure; Data Management; Repository

ACKNOWLEDGMENTS

The repository is developed as part of a three-year project funded by the German Research Foundation from 2013 to 2016 (http://www.radar-projekt.org) and is placed within the program ‘Scientific Library Services and Information Systems (LIS)’ on restructuring the national information services in Germany.

REFERENCES

[1] Wilkinson, M. D. 2016 The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data. 3, 160018 (2016)*. DOI= http://dx.doi.org/10.1038/sdata.2016.18

[2] Treloar, A., Harboe-Ree, C. 2008 Data management and the curation continuum. How the Monash experience is informing repository relationships. In: *14th Victorian Association for Library Automation*, 2008, Conference and Exhibition, Melbourne, Australia. URL= http://arrow.monash.edu.au/hdl/1959.1/43940 (accessed on 04.04.2016)

[3] Kunze et al. (2016) The BagIt File Packaging Format (V0.97). URL= https://tools.ietf.org/html/draft-kunze-bagit-13 (accessed on 05.04.2016)

Establishing Digital Preservation At the University of Melbourne

Jaye Weatherburn
The University of Melbourne
Parkville, Victoria 3010, Australia
jaye.weatherburn@unimelb.edu.au

ABSTRACT

Through 2015-2016, the University of Melbourne is set to achieve the Establishment phase goals for the implementation of digital preservation processes. These goals are detailed in comprehensive Roadmaps, which in turn were developed from the University's Digital Preservation Strategy 2015-2025 [1]. While the Strategy requires implementation across four interrelated areas of digital product, two areas have been prioritized: Research Outputs and Research Data and Records. The phased Roadmaps have been developed to address the challenges inherent to both of these areas. The Roadmaps are comprehensive across organization, culture, policy, and infrastructure, to ensure that the University of Melbourne addresses the challenge of digital preservation of assets through an ongoing commitment to capability building, training, knowledge exchange, advocacy, and ongoing investment in infrastructure (both people and technology). Realizing this vision will support the University's functions, accountability, and legacy.

Keywords

Preservation strategies; Research data; Research outputs; Long-term accessibility; Digital preservation.

1. STRATEGY BACKGROUND

The University of Melbourne's Digital Preservation Strategy articulates a clear vision: to make the University's digital product of enduring value available into the future, thus enabling designated communities to access digital assets of cultural, scholarly, and corporate significance over time.

The Strategy's Establishment phase is being realized during 2015-2016, through clearly developed goals provided by associated Roadmaps. The long-term, ten-year vision of the Strategy requires implementation across four interrelated areas of digital product, in three phases.

The four interrelated areas:

1. Research Outputs
2. Research Data and Records
3. University Records
4. Cultural Collections

The three phases:

Phase 1: Establishment (2015-2016)

Phase 2: Implementation (2016-2017)

Phase 3: Embedding (2017-2025)

The activities stipulated in each phase are underpinned by four key principles around which action is required: Culture, Policy, Infrastructure, and Organization.

2. INSTITUTIONAL CONTEXT

The University of Melbourne aspires to be a public-spirited and internationally engaged institution, highly regarded for making distinctive contributions to society in research and research training, learning and teaching, and knowledge transfer. These aspirations are outlined in the *Growing Esteem* strategy [2].

The University's research strategy, *Research at Melbourne* [3] recognizes the importance of digital assets by declaring that the digital research legacy of the University must be showcased, managed, and preserved into the future. At the same time, open data, open science, and open access initiatives are gaining momentum globally, reflecting changing expectations of government, funding bodies, and the broader community around appropriate access to research product.

Students, researchers, and academics at the University of Melbourne generate and control a considerable number of digital collections of significance. As the number and size of these collections grow, expectations around preservation and access to these digital products – particularly research products – are changing. This is driven by the evolving expectations of researchers, research funding bodies and the broader community. It is within this context that the Digital Preservation Strategy and implementation Roadmaps were endorsed at Academic Board in late 2014.

The Strategy aligns with the University's future-focused motto: *Postera Crescam Laude*: 'We grow in the esteem of future generations' [4], whilst working within the newly implemented, University-wide Melbourne Operating Model (MOM) to realize its goals. The MOM supports a professional services team that aims to provide seamless, easily accessible services to students and academics, so that they can focus on scholarly pursuits.

The MOM encourages innovation, collaboration, and creativity, making for an interesting organizational context to fit digital preservation practice and processes into. Although connected by centralized IT services, various units and departments are still finalizing changes, processes, and ways of operating in the new model, which commenced in 2015. Existing workflows between units must be analyzed, with opportunities for continuous improvement identified, before attempting to introduce digital preservation requirements. An assessment of the qualities, attitudes, and experience levels of staff working in research data and outputs management is also imperative, in order to establish preservation as an integral University process.

The MOM has set the stage for digital preservation to step onto. It acknowledges that every staff member brings a unique set of skills to their work, while increasingly merging and combining these skills into a shared set of seamless processes. The emphasis on shared values for service staff, imparted in various training programs run by the University, will aid the establishment of the Digital Preservation Strategy by encouraging autonomous, creative contributions from skilled workers.

It is becoming evident that the project team members starting the Establishment phase have a vital role to immediately begin a comprehensive, wide-ranging advocacy campaign to many different stakeholders, using jargon-free, discipline-specific examples to illustrate the importance of preservation. Instead of digital preservation being perceived as an extra bureaucratic and financial burden or an 'IT system', the challenge is to present it as a useful tool for future branding and profiling for academics, and also for long-term sustainability of their important research. To emphasize the importance of this component, the Roadmaps prioritize the Organization and Culture aspects of this Establishment phase, before any work is done to determine the more tangible technological infrastructure solutions.

3. PRIORITY ROADMAPS

3.1 Research Outputs and Research Data

While an awareness of the issues, challenges, and requirements for managing and preserving University Records and Cultural Collections must be included in the planning stages of this project, for the purposes of the first Establishment phase in 2015-2016, the Research Outputs and Research Data and Records Roadmaps [5] remain the priority. These two areas have been prioritized due to changing expectations of research funders, government, and the broader community around access to research data and records, and the need to establish a centralized repository for research outputs at the University. Also of high priority is the need to develop and implement consolidated workflows for research data management. It is imperative that the University of Melbourne begins establishing robust infrastructure for managing research data to prepare for a time when funding bodies may mandate and enforce that research data be made openly available.

3.2 Roadmap Goals and Actions

Actions have been identified for addressing each of the Roadmap goals. For the Research Outputs Roadmap, the first phase involves preservation of research outputs using current University infrastructure, and the second phase focuses on the role of the University's planned digital archiving service for preserving research outputs, and also its relationship with the institutional repository. These two phases will be undertaken iteratively, requiring individual business cases for requesting funding. Progress will be monitored against the University's Digital Preservation Strategy, as well as the closely related Roadmap for Research Data and Records.

The Research Data and Records Roadmap is organized into three phases, which will also be undertaken iteratively. Actions include:

- Development of an engagement plan and training framework that articulates the transition in knowledge from research data management to digital preservation
- Implementation of a digital research data repository and digital archiving service underpinned by policy and standards to facilitate preservation of data
- Reviewing and aligning University policies, workflows, and processes related to the management and preservation of research data and records
- Consolidating and coordinating the University's services for supporting the management and preservation of research data and records

4. FROM STRATEGY TO ACTION

Over the first twelve months of establishing the Strategy through the priority Roadmaps, the objectives include:

- Establishing a Panel of Knowledge Experts to guide an engagement plan to educate and advocate University-wide about the benefits and importance of preservation through targeted communications and training strategies, both across the professional staff services area, as well as across academic faculties
- Reviewing and consolidating current digital processes and workflows with regard to management of research outputs and data
- Reviewing and updating University policies, after gaining agreement on the proposed digital processes and workflows to be implemented for preservation
- Developing functional and non-functional requirements for the implementation of a preservation platform

In addition to these goals, an environmental scan of large national and international research institutions engaged in digital preservation projects will be conducted in order to relate relevant elements of their work to the institutional context of the University of Melbourne. It is hoped that this environmental scan will be of use to the wider digital preservation community.

5. ACKNOWLEDGMENTS

My thanks to key contributor Leo Konstantelos in his role as Manager, Digital Scholarship, and Donna McRostie, Associate Director, Research Information Management, both at the University of Melbourne. Many thanks also to Adjunct Professor, Ross Harvey, RMIT University, Melbourne, Australia.

6. REFERENCES

[1] The University of Melbourne Digital Preservation Strategy 2015-2025 (Vision Mandate and Principles) https://minerva-access.unimelb.edu.au/handle/11343/45135

[2] The University of Melbourne Strategic Plan 2015-2020 https://about.unimelb.edu.au/__data/assets/pdf_file/0006/1462065/11364-GROWING-ESTEEM-2015-FA-WEB.pdf

[3] Research at Melbourne: Ensuring excellence and impact to 2025 http://research.unimelb.edu.au/__data/assets/pdf_file/0011/1665722/MelbUniResearchVision_Apr2013.pdf

[4] The University of Melbourne http://about.unimelb.edu.au/tradition-of-excellence

[5] University of Melbourne Digital Preservation Strategy 2015-2025 Implementation Roadmaps https://minerva-access.unimelb.edu.au/handle/11343/45136

Exit Strategies and Techniques for Cloud-based Preservation Services

Matthew Addis
Arkivum
R21 Langley Park Way
Chippenham, UK, SN15 1GE
+44 (0) 1249 405060
matthew.addis@arkivum.com

ABSTRACT

This poster presents an exit strategy for when organisations use cloud-based preservation services. We examine at a practical level what is involved in migrating to or from a cloud-hosted service, either to bring preservation in-house or to move to another service provider. Using work by Arkivum on providing Archivematica as a hosted service, we present how an organisation can use such a hosted service with assurance that they can exit without loss of data or preservation capability. Contractual agreements, data escrow, open source software licensing, use of independent third-party providers, and tested processes and procedures all come into play. These are necessary to mitigate the risks of a wide range of scenarios including vendor failure, service unavailability, change in customer preservation scale or budgets, and migration to or from an in-house approach. There is an existing body of work on how to trust and measure a service that a vendor might offer, for example using audit criteria for Trusted Digital Repositories or measuring service maturity using NDSA preservation levels. However, there has been far less work on how to quickly and safely exit providers of such services - despite this being an essential part of business continuity and disaster recovery. This poster presents some of the considerations and the practical approach taken by Arkivum to this problem including: use of open source software (Archivematica, and ownCloud), data escrow, contracts and handovers, use of vendor independent standards and interfaces (PREMIS, METS, Bagit) and technical migration support, e.g. exports of databases, configurations, software versions and updates. We believe the experience and approach that we have developed will be of use to others when considering either the construction or the use of cloud preservation services.

Keywords

Exit strategy; exit plan; digital preservation; cloud services; software as a service; escrow; migration; hosted preservation.

1. MOTIVATION

Paul Wheatley, Head of Research and Practice at the Digital Preservation Coalition (DPC), presented some of the needs and challenges faced by the DPC membership as part of a talk at PASIG in March 2016 [1]. He articulated that whilst DPC members could see the value of cloud-based preservation services, there were also concerns and barriers to overcome. The top two issues are (a) the need for there to be some form of exit strategy when using a cloud preservation service, and (b) the need for customers of such services to be able to establish trust and perform checks on the quality of the service. Both prevent organisations from adopting preservation services and consequently from achieving the benefits that using these services can offer. This is a problem for the growing number of hosted preservation services, with examples including: Preservica[1], Council of Prairie and Pacific University Libraries (COPPUL)[2], Ontario Council of University Libraries (OCUL)[3], Archivematica hosting and integration with DuraCloud (ArchivesDirect)[4], and Archivematica hosting and integration with Arkivum's archive service in the UK (Arkivum/Perpetua)[5]. Work has been done on the benefits of such cloud services, how to compare and evaluate them, and why exit strategies are important [4][5][6]. These guidelines and comparisons are often based on criteria such as the Data Seal of Approval (DSA)[6], the Trusted Digital Repository standard (ISO16363)[7] or NDSA levels of digital preservation [3]. but don't go into detail on how to implement or verify an exit strategy.

2. APPROACH

Arkivum provides Archivematica[8] as a cloud hosted service, which is integrated with Arkivum's data archiving service. The service includes ownCloud[9] to provide easy upload and download of data. Our approach to providing a built-in exit-strategy for the service's users is to support migration from the Archivematica/Arkivum hosted solution to another Archivematica environment, which might be in-house or might be provided by another service provider. The concept of being able to migrate between preservation environments has been investigated by the SHAMAN [2] project amongst others, but we believe full support for migrating between preservation environments has yet to be implemented in a production preservation service. Given that Archivematica is already open source and supports open specifications (METS, PREMIS, Bagit) then we take the simple case of supporting migration between Archivematica instances rather than the general case of migrating to/from an arbitrary preservation environment. This allows the approach to be simpler and most importantly to be directly tested by users of the service. The approach consists of the following elements.

- All data produced by Archivematica (AIPs and DIPs) are stored in the Arkivum bit preservation service, which includes data escrow. Data escrow consists of a full copy of the data stored with an independent third-party without lock-in to Arkivum. If the user exits the service then there is a contractual agreement that allows them to retrieve the escrowed data directly from the escrow provider. Data is stored on LTO tape using LTFS as a file system. Data is contained within Bagit containers, which provide a manifest of all data files and their checksums. Each file is optionally encrypted using open standards (RSA and AES) and can be decrypted using open source tools if necessary, e.g. OpenSSL, by the user supplying the keys. Each data file is accompanied by an XML 'sidecar' file that contains metadata on the file, e.g. when it was originally ingested, which encryption keys were used, and the original file name, path and attributes. In this way, the user can retrieve their AIP and DIP files without lock-in.
- Archivematica databases and configuration are exportable from the service and can be downloaded by the user on a regular basis. For example, this includes Archivematica's internal database for storing processing state and AIP/DIP information, webserver configuration (nginx), indexes made of the files processed by Archivematica (elasticsearch). This export allows the user to in effect 'backup' their hosted Archivematica pipeline and storage service. The databases and configurations are snapshotted on a regular basis. This allows the ongoing 'state' of the service to be recorded and replicated into the users' environments.
- Log files are provided of the software versions and updates used in the hosted service, e.g. version of the Archivematica pipeline and storage service, underlying operating system, and peripheral services such as ownCloud. These logs are exported to allow the user to create their own record of the software versions used in the hosted service. This ensures that if the users try to recreate the service then they can do so using the same software versions and hence will be able to import/overlay the database and configuration backups.
- The database and configuration backups along with software version and update logs are all exported through ownCloud. This allows the user to automatically synchronise a local copy of these files into their environment without the need to explicitly remember to download them on a regular basis. Along with the AIP and DIPs stored in data escrow this means that the user has access to both their data and the information needed to take this data and rebuild a working Archivematica system around it.

We are currently working on a simple way for users to do a 'migration test' to verify that the information and data described above is complete and sufficient. Whilst it is easy to assert to a user that everything necessary has been done, the best way to validate this in practice is to perform an actual migration and demonstrate that a working Archivematica instance can be built from the supplied inputs. Arkivum already does this for data escrow through a 'USB key' based escrow test. When using the bit-preservation service the user can specify a test dataset that they want to use for an escrow test. This test data set is 'escrowed' to a USB key and delivered straight to the customer (or via the escrow provider if desired). The user can then validate that the escrowed data is recoverable and is identical to the test data set that they supplied. We are developing a similar approach for Archivematica. The user will be able to set up and use a 'test pipeline' in the hosted service and then ask for this to form the basis of a 'migration test'. The database and configuration etc. for this pipeline will be exported along with the test AIPs and DIPs that it generates. In a similar way to the escrow test, this will be delivered to the user in a self-contained form, e.g. USB key. We aim for this to include a working Archivematica instance configured using the test dataset and exports from the service, for example provided as a bootable drive. In this way, the user will be able to compare and validate that the migration successfully replicates the test pipeline in the hosted service. The test helps provide assurance that the full production pipelines can also be migrated if/when needed. This is important because the production pipelines may contain substantial amounts of data and hence doing actual migration tests of the whole service on a regular basis will typically not be practical.

3. CONCLUSION

Hosted preservation services offer many benefits but their adoption can be hampered by concerns over vendor lock-in and inability to migrate away from the service, i.e. lack of exit-plan. We have used Archivematica as a hosted service to investigate what is needed in practice to migrate from the service to an independent Archivematica instance. The approach includes data escrow, export of state information (e.g. databases and configuration), and most importantly a way for users to independently test and verify that migration is possible, i.e. the exit strategy can be successfully executed in practice.

4. REFERENCES

[1] Wheatley, Paul. 2016. The DPC Community: Growth, Progress and Future Challenges). Preservation and Archiving Special Interest Group meeting (PASIG), Prague. http://pasig.schk.sk/wordpress/agenda

[2] Watry, Paul. 2007. Digital Preservation Theory and Application: Transcontinental Persistent Archives Testbed Activity. The International Journal of Digital Curation. Issue 2, Volume 2. www.ijdc.net/index.php/ijdc/article/download/43/28

[3] Peltzman, Shira. 2016. Expanding NDSA Levels of Preservation. http://blogs.loc.gov/digitalpreservation/2016/04/expanding-ndsa-levels-of-preservation/

[4] Beagrie, Neil. 2015. Guidance on Cloud Storage and Digital Preservation. How Cloud Storage can address the needs of public archives in the UK. Report from The National Archives. http://www.nationalarchives.gov.uk/documents/CloudStorage-Guidance_March-2015.pdf

[5] AVPreserve. 2014. Cloud Storage Vendor Profiles. https://www.avpreserve.com/papers-and-presentations/cloud-storage-vendor-profiles/

[6] Digital Preservation Handbook, 2nd Edition, Digital Preservation Coalition © 2015 http://www.dpconline.org/advice/preservationhandbook

[1] http://preservica.com/

[2] http://www.coppul.ca/archivematica

[3] http://www.ocul.on.ca/node/4316

[4] http://duracloud.org/archivematica

[5] http://arkivum.com/blog/perpetua-digital-preservation/

[6] http://datasealofapproval.org/en/

[7] http://public.ccsds.org/publications/archive/652x0m1.pdf

[8] https://www.archivematica.org/en/

[9] https://owncloud.org/

Geospatial Data: Conservation and Archiving Planning

Helen Gollin
Federal Office of Topography
swisstopo
Seftigenstrasse 264
3084 Wabern, Switzerland
+41 58 469 04 40
helen.gollin@swisstopo.ch

Urs Gerber
Federal Office of Topography
swisstopo
Seftigenstrasse 264
3084 Wabern, Switzerland
+41 58 469 02 82
urs.gerber@swisstopo.ch

Martin Schlatter
Federal Office of Topography
swisstopo
Seftigenstrasse 264
3084 Wabern, Switzerland
+41 58 469 03 04
martin.schlatter@swisstopo.ch

ABSTRACT

In this poster, we describe a work package of the project Ellipse - archiving of official geospatial data under federal legislation in Switzerland. The work package treats the Conservation and Archiving Planning for all the geospatial data of the federal administration. The Conservation and Archiving Planning allows to determine how long a set of geospatial data is kept at the authority responsible and if it is of archival value or not. An overarching, coordinated and joint planning is of fundamental importance when dealing with geospatial data sets, so to ensure the combinability in the long term.

Keywords

Long-term preservation, preservation planning, conservation and archiving planning, long-term availability, archiving, geospatial data, OAIS, metadata, management of geospatial data, appraisal of geospatial data.

1. INTRODUCTION

The project Ellipse is carried out as a joint project involving the Swiss Federal Office of Topography (swisstopo) and the Swiss Federal Archives. Its aim is to find a solution for the archiving of geospatial data in the federal administration. A set of objective stated:

- The solution should be developed for the entire federal administration.
- It should be a well-founded, integral solution for long-term availability and archiving.
- It must permit archived digital geospatial data to be (subsequently) re-integrated into a geographic information system (GIS). It must enable geoinformation to be restored at a later date.

To ensure this well-founded and integral solution for long-term availability and archiving the work package Conservation and Archiving Planning (CAP) was realised. More information, especially to other aspects of the archiving of geospatial data can be found in the concept report of the project Ellipse [1].

2. CONCEPTION OF THE CONSERVATION AND ARCHIVING PLANNING

What is to happen to the geospatial data in future, in other words which data are to be available where, for how long, and for what purpose, is a key issue in the management of geospatial data. In Switzerland there is a legal framework for answering these questions, which distinguishes between *conservation* for a limited time at the authority responsible (long-term availability) and *archiving* for an unlimited time by the Swiss Federal Archives.

The archival value of all documents must be assessed before archiving. To do this, the bodies required to offer records for safekeeping and the Swiss Federal Archives assess corporately which of the documents offered are worth archiving, and which should be destroyed once their conservation period has ended.

The Swiss Federal Archives operate a standard method for appraising documents against a catalogue of criteria which is applied equally to all types of documents. The criteria and the two-stage overall appraisal process can also be applied to geospatial data. In view of the important interdependencies between the geospatial data collected by various authorities, the procedure has been supplemented such that, when appraising in accordance with legal and administrative criteria, not only the authority responsible for the data according to the law but also, via the latter, other responsible authorities that are affected, are involved.

The aim of long-term availability is to conserve official geospatial data for a limited period in such a way that their quantity and quality are maintained and they are available for continuous active use. Online availability should extend not just to the data that are current at a given time but also to defined older versions (in the sense of time series) to enable amongst others monitoring

The archive and the authorities responsible must draw up an overarching, coordinated and joint conservation and archiving plan. Appraisal of geospatial data for time-limited conservation in long-term availability and subsequent archiving, where appropriate, are to be planned and coordinated in advance and not on a case-by-case basis, if questions of appraisal of an individual geospatial data set are upcoming.

Although the goals and statutory basis of long-term availability and archiving differ, they nevertheless relate to the same documents (in this case geospatial data) and require detailed reflection on their function, potential use and links, as well as the exploitation of possible synergies. Linking the two decision-making processes together from an organisational point of view is therefore a matter of importance.

To maximise the benefit from the potential synergies between the selection of geospatial data for long-term availability and appraisal for archiving, coordination is advisable on two levels: coupling the two processes together; and applying them to all federal geospatial data sets. The advantages of this approach are as follows:

- First, linking the prospective appraisal of all federal geospatial data with regard to long-term availability and archiving enables the two aspects of limited conservation and (unlimited) archiving to be coordinated.
- Second, registration of all geospatial data on a single occasion creates a shared working basis, which is preferable to individual stocktakes in terms of both the work involved and the information value.
- Third, early planning for long-term availability and archiving enables the various parties involved to input their requirements and interests into the process.
- Fourth, account can be taken of the interdependencies between thematic geospatial data and geospatial reference data or geospatial data. As all parties are involved at the same time, the results can be aligned where necessary.
- Fifth, coordination takes account of the fact that geospatial data, the vast majority of which are collected decentrally, can be linked to geoinformation in any number of ways. This needs to be borne in mind both in long-term availability and in the archive.
- Sixth, the workload involved at a later stage when geospatial data are submitted to the archive is significantly reduced.

In addition to efficiency gains, this approach therefore permits a holistic perspective on the issue of what is to happen to the various geospatial data. If transparency is assured and an overall view is available on this point, geospatial data can be managed prospectively and their long-term usability is secured. Geospatial data that are no longer needed can be filtered out at an early stage, instead of unnecessarily consuming resources. Finally, planning is a prerequisite for the automation of transfer between geospatial data-producing authorities and the archive. It also creates transparency for all involved as well as for users.

3. REALISATION OF THE CONSERVATION AND ARCHIVING PLANNING

The conception of the Conservation and Archiving Planning was done in 2013. From 2014 till 2016 the Conservation and Archiving Planning was realised with all the federal offices involved.

Initially, an inventory was generated, this included on the one hand side the compilation of the official geospatial data under federal legislation and on the other hand side other geospatial data that the federal offices produce. The inventory was generated in a tool, which was also used by all the participants to fill in their appraisal information. When the inventory was completed, the authority responsible appraised the long-term availability and the archival value of their geospatial data sets. In the implementation this means they defined how long their geospatial data sets are going to be kept in the long-term availability and if their geospatial data sets are of archival value or not from a legal and administrative point of view. When every responsible authority had finished the appraisal of their geospatial data sets, the other authorities had the opportunity to appraise these geospatial data sets as well. For this purpose a workshop was conducted, where all the geospatial data producing offices, respectively the authorities responsible, came together and could place their requirements. The main part of the changes due to this workshop where harmonisations of the appraisal of geospatial data sets with a similar data model but diverging authorities responsible (e.g. sectoral plans). After this workshop the Swiss Federal Archives conducted the appraisal of the archival value of all the geospatial data sets from a historical and social point of view. This process was finalized with an official appraisal decision from the direction of the Swiss Federal Archives.

4. CONCLUSIONS

The activities of the initial Conservation and Archiving Planning were completed in spring 2016. They are published on the geoportal of the Swiss Confederation[1] and on the website of the Swiss Federal Archives[2]. By the end of 2016 the annual updating process is going to be designed, so that from 2017 onwards the Conservation and Archiving Planning can be put into operation. The annual updating process renders the possibility to adjust the appraisal when new geospatial data sets are generated or when conditions change.

5. REFERENCES

[1] Project Ellipse, 2013. *Concept for the archiving of official geodata under federal legislation*. Concept report. Federal Office of Topography swisstopo, Swiss Federal Archives, Berne, Switzerland. http://www.swisstopo.admin.ch/internet/swisstopo/en/home/topics/geodata/geoarchive.parsysrelated1.59693.downloadList.56737.DownloadFile.tmp/konzeptberichtellipsev1.3publikationen.pdf

[1] http://www.geo.admin.ch/internet/geoportal/de/home/topics/archive_planning.html

[2] https://www.bar.admin.ch/dam/bar/de/dokumente/bewertungsentscheide/Geobasisdaten%20Bewertungsentscheid%202016.pdf.download.pdf/Bewertungsentscheid%20Geo(basis)daten%20des%20Bundes%20(Projekt%20Ellipse,%20AAP),%202016-02-19.pdf

HydraDAM2: Building Out Preservation at Scale in Hydra

Heidi Dowding
Indiana University
1320 E. 10th St
Bloomington, IN 47401
+1 (812) 856-5295
heidowdi@indiana.edu

Michael Muraszko
WGBH
1 Guest St
Boston, MA 02135
+1 (617) 300-2000
michael_muraszko@wgbh.org

ABSTRACT

HydraDAM2, a digital preservation repository project being developed through a partnership between Indiana University and WGBH, aims to leverage new developments in the Hydra/Fedora stack in order to provide better long-term management solutions for large audiovisual files.

Keywords

Hydra; Fedora; Repositories; Digital Preservation; Audiovisual

1. INTRODUCTION AND BACKGROUND

Indiana University and WGBH, a large academic research institution and a public media organization respectively, are currently both managing large audiovisual collections. While WGBH regularly manages multimedia as part of its daily production, IU's current developments in this area are based on the Media Digitization and Preservation Initiative (MDPI), which will result in 6.5TB of digital audiovisual content. [1] As both institutions have identified similar challenges in managing large- scale audiovisual content, Indiana University and WGBH have partnered to develop a repository aimed at long-term management and preservation.

This repository project, titled HydraDAM2, will build on WGBH's original NEH-funded Hydra Digital Asset Management system (HydraDAM) [2] as well as IU's IMLS-funded Avalon Media Systems. [3] The original HydraDAM is a digital file preservation repository built for the long-term storage of media collections. Like many Hydra applications, HydraDAM is a web- based, self-deposit system that allows for the search and discovery of the files and records stored in the repository. Storage for the HydraDAM repository is limited to the server or virtual machine on which the application is installed. Avalon Media Systems is a Hydra digital access repository aimed at discoverability and use of audiovisual materials.

HydraDAM2 will leverage recent improvements to the Fedora repository. The new digital preservation repository will allow for the storage of files either online via a Hierarchical Storage Management (HSM) system or offline via LTO data tape or hard drives. Having begun work in mid-2015, the HydraDAM2 team will complete a minimum viable product to be implemented within each institution by the fall of 2016. The ultimate goal of HydraDAM2 is to create an extensible product that can be reused within any Hydra institution.

2. IDENTIFIED GAPS IN HYDRA

One of the main limitations of the current Hydra / Fedora technology stack identified by the HydraDAM2 team is the inability to store large digital files within Fedora. This has been challenging with web-based repositories because there are often limits on size when ingesting files into Fedora from a web browser. Processing large files for things like fixity and characterization is also problematic, as it can be difficult to pinpoint the problem if any processes get held up or fail.

Another identified challenge in Hydra is the favoring of self-deposit systems where a majority of the metadata describing an object is generated during the ingest process. This is a problem for many institutions dealing with years of metadata records, sometimes from legacy digital asset management systems. In moving to a new Hydra self-deposit system, an institution could immediately have a significant backlog of files that would require re-description upon ingest. Hydra self-deposit repository systems are most successful for new projects, not for migration of legacy files and metadata.

3. HYDRADAM2 STACK

The HydraDAM2 system is based on the open source Hydra repository application framework and will utilize the emerging Fedora 4.0 digital repository architecture. There has also been a recent development in data modeling in Fedora. The Portland Common Data Model (PCDM) is a flexible, extensible domain model that is intended to underlie a wide array of repository and DAM applications. By implementing PCDM in HydraDAM2, we hope that using an emerging, standardized model for our data will allow for better understanding and interoperability with current and future Hydra open source solutions.

4. MAJOR FUNCTIONALITY

4.1 Management of Large Files

One of the main aims of the HydraDAM2 project is to reconcile the challenge of large files within the Hydra/Fedora environment by building out mechanisms for connection between local storage architectures and the HydraDAM2 repository. At Indiana University, this will likely integrate an API developed for asynchronous interactions with the institution's HSM storage backend utilizing Apache Camel routes as a means of integration. This scenario will allow for better management of terabyte-size audiovisual files within HydraDAM2, as the content will be safely deposited in IU's storage backend but manageable through HydraDAM2. The implementation at WGBH will be somewhat simpler in allowing HydraDAM2 to interact with their LTO tape storage backend.

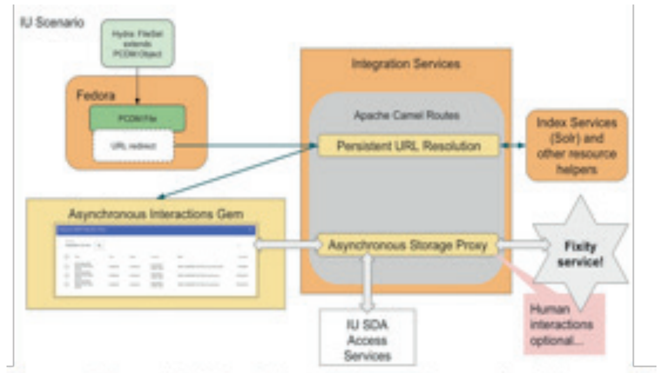

IU Scenario
Hydra FileSet extends PCDM Object
Fedora
PCDM File
URL redirect
Integration Services
Apache Camel Routes
Persistent URL Resolution
Index Services (Solr) and other resource helpers
Asynchronous Interactions Gem
Asynchronous Storage Proxy
Fixity service!
IU SDA Access Services
Human interactions optional...

Figure 1. Indiana University Storage Interactions [4]

4.2 Reporting and Legacy Technical Metadata Management

Another goal of HydraDAM2 is to build out preservation functionality within Hydra and make it reusable. A majority of this functionality is focused on generating reports. Utilizing search functionality from the Blacklight piece of Hydra, HydraDAM2 expands capabilities in working with technical metadata for discoverability and management. This will result in the end user's ability to generate reports on things like file format, date created, and fixity events. HydraDAM2 will also include the ability for users to ingest previously created technical metadata so the system does not have to process files on ingest and generate them. As both institutions are managing collections with significant amounts of legacy metadata, this feature is crucial to scaling the repository solution.

4.3 Ongoing Curation

The final overarching goal of HydraDAM2 is to create an environment for ongoing management of digital files. Where Avalon will function as the access repository for all of IU's audiovisual content, HydraDAM2 will provide mechanisms for preservation and sustainability of content. While the first iteration of the repository focuses on basic preservation events like scheduled and triggered fixity checks, future iterations could include functionality like pathways for format migration. The main aim is to create a reusable Hydra repository with functionality for the necessary ongoing preservation functions related to audiovisual content.

5. CONCLUSION

As "an ecosystem of components" aimed at allowing individual institutions to more efficiently and effectively meet their repository needs, the Hydra project is constantly identifying gaps in infrastructures and workflows. As part of this, the HydraDAM2 digital preservation repository will fill in the gaps identified in the ongoing curation and management of large audiovisual files. By jointly developing this repository as a partnership between two very disparate institutions with two diverse storage backends, the end result will be a new set of functionality that can be utilized at a broad variety of institutions.

6. ACKNOWLEDGMENTS

This project has been developed based on a grant from the National Endowment for the Humanities. Our thanks to all staff involved in the project at Indiana University and WGBH.

7. REFERENCES

[1] https://mdpi.iu.edu/

[2] https://github.com/projecthydra-labs/hydradam

[3] http://www.avalonmediasystem.org/

[4] Floyd, R. (February 22, 2016). HydraDAM at IU. Presented at HydraDAM2 Partners Meeting: Bloomington, IN.

Project EMiL – Emulation of Multimedia Objects

Tobias Steinke
Deutsche Nationalbibliothek
Adickesallee 1
60322 Frankfurt am Main, Germany
t.steinke@dnb.de

Frank Padberg
Karlsruhe University of Art and Design
Lorenzstrasse 15
76135 Karlsruhe, Germany
fpadberg@hfg-karlsruhe.de

Astrid Schoger
Bayerische Staatsbibliothek
Ludwigstraße 16
80328 München, Germany
astrid.schoger@bsb-muenchen.de

Klaus Rechert
University of Freiburg
Hermann-Herder Str. 10
79104 Freiburg, Germany
klaus.rechert@rz.uni-freiburg.de

ABSTRACT

In this poster we will present the results of the German research project EMiL (Emulation of Multimedia objects in Libraries). In the project, an emulation-based flexible and automatable access framework for multimedia objects in libraries and museums was developed.

Keywords

preservation strategies; emulation; digital libraries; multimedia; emulation frameworks.

1. INTRODUCTION

The German project EMiL (Emulation of Multimedia objects in Libraries)[1] focused on providing access to multimedia objects of the 90s that usually are stored on CD-ROMs [7]. Objects such as digital encyclopedias or interactive art pieces are difficult to present using current access systems in a reading room or exhibition space, since their original, now out-dated run time environments typically are not supported by the existing access system. This calls for an emulation of the vintage run time systems in order to provide access [5], similar to other research projects on access frameworks [1][2][3][6].

There are several technical challenges for an access framework based on emulation. Because of the variety of possible objects and environments, the EMiL framework must be able to employ a range of different emulators, existing and future ones. Given the huge size of digital collections, especially in libraries, automated procedures are needed. The EMiL framework includes a tool that automatically identifies an emulation environment that supports a chosen object [4]. After the (automated or manual) selection of the run time environment, the EMiL framework deploys the emulation, executes the object in the emulated environment, and shuts down the emulation after usage. Access to the emulated multimedia object is provided to the onsite user through a web browser interface.

EMiL aims at integrating with different catalogues and long-term archiving systems. While the integration with library catalogue systems is work in progress, the integration with a particular long-term archive system has already been tested.

2. POSTER CONTENT

The poster will describe the goals and challenges addressed in the EMiL project. It will visualize the components of the framework and its embedding into the existing infrastructures in libraries and museums.

Sample screenshots of multimedia objects in emulation will give an impression of the access interface (see Figure 1 for a sample cutout).

The poster will also describe reuse possibilities of the EMiL framework.

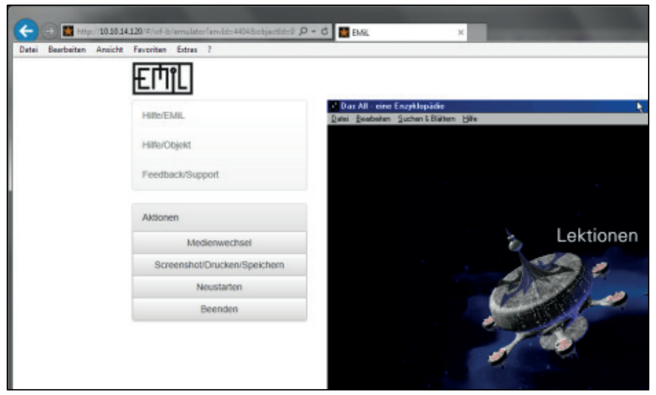


Figure 1. The EMiL access interface (cutout)

3. ACKNOWLEDGMENTS

The project is funded by the German Research Foundation (DFG).

4. REFERENCES

[1] Braud, M., Lohman, B., and van der Hoeven, J. 2012. How to run emulators remotely via the Emulation Framework. Online at http://emuframework.sourceforge.net/docs/EF-howto-remoteemulation-1.0.pdf

[2] Farquhar, A., and Hockx-Yu, H. 2007. Planets: Integrated Services for Digital Preservation. *Int. Journal of Digital Curation* IJDC 2, 2 (2007), 88-99.

[3] Holdsworth, D., and Wheatley, P. 2001. Emulation, Preservation and Abstraction. In *Research Libraries Group RLG DigiNews* 5, 4 (Aug. 15, 2001) Online at http://sw.ccs.bcs.org/CAMiLEON/dh/ep5.html

[4] Rechert, K., Liebetraut, T., Stobbe, O., Valizada, I. and Steinke, T. 2015. Characterization of CD-ROMs for Emulation-Based Access. *Proceedings of the 12th International Conference on Digital Preservation (iPRES 2015)*, p. 144 – 151.

[5] Rothenberg, J. 1995. Ensuring the Longevity of Digital Information. *Scientific American*, 272(1) (Jan. 1995), 42-47

[6] Satyanarayanan, M. 2013. Olive: One-click Execution of Internet-Archived Software. In *NYU Scientific Reproducibility Workshop* (New York, USA, May 30, 2013).

[7] Steinke, T., Padberg, F., and Schoger, A. 2015. Project EMiL: Emulation-based Access Framework. *D-Lib Magazine.* Volume 21, Number 9/10, http://dlib.org/dlib/september15/09inbrief.html.

[1] http://www.multimedia-emulation.de/

Should We Keep Everything Forever?: Determining Long-Term Value of Research Data

Bethany Anderson
University Archives
University of Illinois at Urbana-Champaign
bgandrsn@illinois.edu

Susan Braxton
Funk Library
University of Illinois at Urbana-Champaign
braxton@illinois.edu

Elise Dunham
Research Data Service
University of Illinois at Urbana-Champaign
emdunham@illinois.edu

Heidi Imker
Research Data Service
University of Illinois at Urbana-Champaign
imker@illinois.edu

Kyle Rimkus
Preservation Unit
University of Illinois at Urbana-Champaign
rimkus@illinois.edu

ABSTRACT

The University of Illinois at Urbana-Champaign's library-based Research Data Service (RDS) launched an institutional data repository called the Illinois Data Bank (IDB) in May 2016. The RDS makes a commitment to preserving and facilitating access to published research datasets for a minimum of five years after the date of publication in the Illinois Data Bank. The RDS has developed guidelines and processes for reviewing published datasets after their five-year commitment ends to determine whether to retain, deaccession, or dedicate more stewardship resources to datasets. In this poster, we will describe how the University of Illinois at Urbana-Champaign preservation review planning team drew upon appraisal and reappraisal theory and practices from the archives community to develop preservation review processes and guidelines for datasets published in the Illinois Data Bank.

Keywords

Innovative practice; appraisal; digital preservation; archival theory

1. INTRODUCTION

The Illinois Data Bank's [4] purpose is to provide University of Illinois at Urbana-Champaign researchers with a library-based repository for research data that will facilitate data sharing and ensure reliable stewardship of published data. The initiating goal that the IDB fulfills is that it provides a mechanism for researchers to be compliant with funder and/or journal requirements to make results of research publicly available. More broadly, the IDB endeavors to promote the discoverability and use of open research data by offering a preservation and access solution that is trusted by researchers at the University of Illinois at Urbana-Champaign. The Research Data Service (RDS) currently commits to preserve data and make it available for at least five years from the date of publication in the IDB.

In order to ensure that we are able to fulfill our commitment to stewarding University of Illinois at Urbana-Champaign research data in an effective and scalable manner, the RDS has established a policy framework that enables us to assess the long-term viability of a dataset deposited into the IDB. The RDS has developed guidelines and processes for reviewing published datasets after their five-year commitment ends to determine whether to retain, deaccession, or dedicate more stewardship resources to datasets. Enacting a systematic approach to appraising the long-term value of research data will enable the RDS to allot resources to datasets in a way that is proportional to the datasets' value to research communities and its preservation viability.

2. PRESERVATION REVIEW

In this poster we will present the preservation review guidelines and processes we have developed within the context of archival appraisal theory and practice [1][2][3][5]. We will describe the automated measures we will implement to prioritize datasets for preservation review, as well as outline the Preservation Review Guidelines that preservation "Assessment Teams" will use to determine whether to retain, deaccession, or dedicate more stewardship resources toward datasets that undergo preservation review. The poster will also demonstrate the intended personnel make-up of "Assessment Teams" and examples of how dataset disposition will be documented and presented to IDB users.

The Illinois Data Bank Preservation Review Guidelines, which will be featured and expanded upon in this poster, are given in Table 1.

3. REFERENCES

[1] Haas, J.K., Samuels, H.W. and Simmons, B.T. 1985. Appraising the records of modern science and technology: a guide. Massachusetts Institute of Technology.

[2] Society of American Archivists, Technical Subcommittee on Guidelines for Reappraisal and Deaccessioning (TS-GRD). 2012. Guidelines for Reappraisal and Deaccessioning http://www2.archivists.org/sites/all/files/GuidelinesForReappraisalAndDeaccessioning-May2012.pdf.

[3] UK Data Service. 2014. Collections Development Selection and Appraisal Criteria version 01.00 https://www.ukdataservice.ac.uk/media/455175/cd234-collections-appraisal.pdf

[4] University of Illinois, Research Data Service. 2016. Illinois Data Bank https://databank.illinois.edu.

[5] Whyte, A. and Wilson, A. 2010. How to Appraise and Select Research Data for Curation. DCC How-to Guides. Edinburgh: Digital Curation Centre http://www.dcc.ac.uk/resources/how-guides

Table 1. Preservation review guidelines for the Illinois Data Bank

Evaluated by Curators/Librarians/Archivists

Criterion	*Consideration*
Cost to Store	What is the estimated cost of continuing to store the dataset?
Cost to Preserve	What is the estimated cost of continuing or escalating preservation for the dataset? Preservation actions may include file format migration, software emulation, and/or enhancement of preservation metadata.
Access	What do download and page view metrics indicate about interest in this dataset over time?
Citations	Has the dataset been cited in any publications?
Associated Publication Citations	If the dataset supports the conclusions of a publication, has that publication been cited in any other publications?
Restrictions	Does the dataset have any access or re-use restrictions associated with it?

Evaluated by Domain Experts

Criterion	*Consideration*
Possibility of Re-creation	Is it possible to create the dataset again?
Cost of Re-creation	If it is possible to create the dataset again, what would be the cost of doing so?
Impact of Study	Did the study that generated this dataset significantly impact one or more research disciplines?
Uniqueness of Study	Was the study that generated this dataset novel?
Quality of Study	Is the study that generated this dataset regarded as being of quality by domain experts?
Quality of Dataset	Is the dataset of quality according to domain experts?
Current Relevance	Is the dataset useful for addressing contemporary research questions according to domain experts?

Evaluated by Curators/Librarians/Archivists and Domain Experts

Criterion	*Consideration*
Availability of Other Copies	Is the copy of the dataset in the Illinois Data Bank the only one?
Understandability	Has the creator supplied sufficient metadata and documentation related to the dataset’s creation, interpretation, and use in order to facilitate future discovery, access, and reuse?
Dependencies	Are the software, computing environment, or other technical requirements for using the dataset known? If so, are they available?
Appropriateness of Repository	Is there another trusted repository that, based on their collecting scope and user community, would be a better home for the dataset?

The Status of Compliance with OAIS Reference Model in the National Diet Library Digital Collections

Tsukasa Kimezawa
Digital Information Services Division,
Digital Information Department, National Diet Library
1-10-1 Nagata-cho, Chiyoda-ku,
Tokyo, Japan 100-8924
(+81) 335063328
t-kimeza@ndl.go.jp

Shuji Kamitsuna
Kansai-Kan,
National Diet Library
8-1-3 Seikadai, Seika-cho,
Soraku-gun, Kyoto, Japan 619-0287
(+81) 774981484
kamituna@ndl.go.jp

ABSTRACT

The National Diet Library (NDL) has been providing access to digitized library materials via the Internet since 2002. The NDL has been digitizing books and magazines continuously since then, and collecting digitized materials from other institutions. In addition to digitized materials, NDL began to collect online publications (electronic books and electronic magazines) that were not protected by Digital Rights Management (DRM) from the Internet in 2013. The NDL Digital Collections was launched in 2011 to collect, preserve, and distribute these materials. This paper provides an overview of the NDL Digital Collections and discusses current achievements as well as the challenges faced in effecting long-term preservation while meeting the functional requirements of the OAIS reference model.

Keywords

Digital preservation, Long-term accessibility, OAIS reference model

1. THE NDL DIGITAL COLLECTIONS

The NDL digitizes and provides access via the Internet for books, magazines, and other library materials. Prior to 2009, digitization of materials had been limited to materials confirmed to be in the public domain as a result of copyright investigation or for which permission had been obtained either from the copyright owners or the Agency of Cultural Affairs. The 2009 revision of the Copyright Law enabled the NDL to digitize books and magazines for preservation even if they were not yet in the public domain and to provide access to them on the premises of the NDL at the Tokyo Main Library, Kansai-kan, or International Library of Children's Literature. That same year, a major budget allocation for digitization enabled the NDL to digitize more than 2 million library materials. (Table 1)

Table 1. Number of digitized materials

Kind of Materials	Number
Periodicals	1,240,000
Books	900,000
Online Publications	300,000
Doctoral Dissertations	140,000
Rare books and Old Materials	90,000
Historical Recordings Collection	50,000
Others	70,000
Total	2,790,000

The original system, however, lacked both scalability of storage and the functionality to stream digitized recordings. In particular, there were several systems in place, which were difficult to operate and maintain individually in terms of cost and manpower.

The NDL Digital Collections (Figure 1) was developed and integrated with existing systems in 2011 to collect and preserve a wide variety of materials. Functionality for providing people with visual disabilities access to Digital Accessible Information SYstem (DAISY) materials was added in 2012, for enabling publishers to upload online publications with metadata to the NDL Digital Collections was added in 2013, and for collecting doctoral dissertations in electronic format as well as for transmitting digitized material to public libraries was added to the system in 2014.

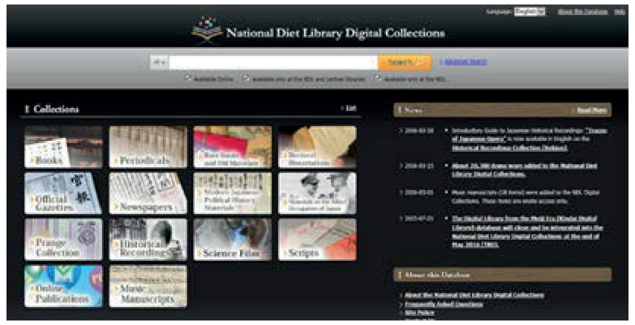


Figure 1. The NDL Digital Collections

The NDL strives to conform to all legal requirements in digitizing books and magazines as well as collecting and preserving online publications from the Internet. It also provides a variety of services to diverse users, who include Diet members and their staff, employees of government agencies, and patrons from the general public, both in Japan and overseas. Digitized materials in the NDL Digital Collections have bibliographic data, which enables even patrons who lack any specialized knowledge to search and access digitized materials through the system's browsing function.

2. COMPARISON WITH OAIS REFERENCE MODEL

We have described how the NDL Digital Collections was designed to collect, preserve, and distribute digitized materials in variety of diverse formats. Next, we will discuss how it compares with the OAIS reference model (ISO 14712: 2012).

The OAIS reference model defines six functional entities: Ingest, Archival Storage, Data Management, Administration, Access and Preservation Planning. Digital information for preservation is handled as an information package in the OAIS reference model.

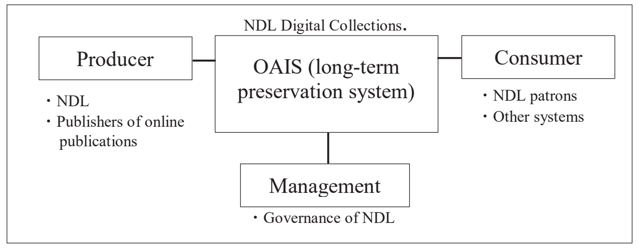


Figure 2. OAIS reference model and NDL Digital

The individual Archival Information Packages (AIP) that comprise the NDL Digital Collections are not preserved in a single comprehensive archival file. Instead, image files of digitized materials are placed in archival storage per each book or magazine. On the other hand, preservation metadata and package information are stored in a Relational DataBase Management System (RDBMS) that is separate from the archival storage. This configuration was adopted in consideration of accessibility to preservation metadata and package information as well as future flexibility to change information package file formats.

The NDL Digital Collections ingests numerous digitized materials by means of a collective registration method. This method directly collects digitized materials from external hard disks. There are also functions for collecting electronic books and electronic magazines via web crawlers or by uploading via a website.

The number, contents, and metadata of ingested materials are checked by staff using an administrative interface. The file format and required metadata items are checked by the system automatically. If necessary, the format of ingested image files is modified to a more suitable format for access and preservation. For example, an image file in TIFF format are converted to JPEG2000.

Preservation metadata is compliant with PREMIS ver. 2.2. Message digest, the date and hour of registration, the staff ID and the file format are recorded as fixed information. Persistent identifiers in info URI are given as reference information. DOIs are given to a part of digitized materials, such as doctoral dissertations in electronic format or rare books and old materials.

In addition to information obtained when materials are digitized, other bibliographic data is added from an integrated search service, NDL Search, for use as metadata when searching. An extended NDL Dublin Core Metadata Description (DC-NDL) is used as the format for metadata information.

We adopted the most suitable devices available for archival storage. In addition to the cost and capacity of devices for archival systems, other factors taken into consideration were creditability, read/write capability, and scalability. Since 2011, we have adopted a distributed file system and constructed a petabyte-class storage system.

Data management of information such as metadata, access restriction, and digital rights is performed by RDBMS.

Administration duties performed by NDL staff include negotiating with publishers and making policies. Monitoring tools are used to administer systems and to collect logs and statistics automatically.

Access management is performed by the NDL Digital Collections server, which converts digital material archived in JPEG2000 format to JPEG and transmits it per each patron request. Audio and video recordings are streamed to the patron from the server. DAISY materials are provided to visually impaired patrons via streaming and download.

3. NEXT STEP

We face a number of future challenges in providing functionality that is not yet implemented within the NDL Digital Collections. In particular, we have yet to implement Preservation Planning to provide environments for accessing to obsolescent file formats. We also need to negotiate with publishers regarding the preservation of online publications with DRM.

Ensuring long-term access to materials that are difficult to migrate involves preserving environments, including hardware and OS, and applications as well as instructions on how to use them. In the future, emulation will be an essential part of the NDL Digital Collections.

4. REFERENCES

[1] National Diet Library, National Diet Library Digital Collections, http://dl.ndl.go.jp/?__lang=en

[2] T. Kimezawa, "The Status of compliance with OAIS reference model in the "National Diet Library Digital Collections" Transformation from "Digital Library from the Meiji Era," *Journal of Information Processing and Management,* vol. 58, pp.683-693, 2015.

[3] PREMIS Editorial Committee, PREMIS Data Dictionary for Preservation Metadata. 2012, Version 2.2, 265p, http://www.loc.gov/standards/premis/v2/premis-2-2.pdf

To Act or Not to Act – Handling File Format Identification Issues in Practice

Matthias Töwe
ETH Zurich, ETH-Bibliothek
Rämistrasse 101
8092 Zurich, Switzerland
+41-(0)44 632 60 32
matthias.toewe@library.ethz.ch

Franziska Geisser
ETH Zurich, ETH-Bibliothek
Rämistrasse 101
8092 Zurich, Switzerland
+41-(0)44 632 35 96
franziska.geisser@library.ethz.ch

Roland E. Suri
ETH Zurich, ETH-Bibliothek
Weinbergstr. 74
8006 Zurich, Switzerland
+41-(0)44 632 39 19
roland.suri@library.ethz.ch

ABSTRACT

Format identification output needs to be assessed within an institutional context, also considering provenance information that is not contained in the data, but provided by data producers by other means. Sometimes, real issues in the data need to be distinguished from warnings. Ideally, this assessment should permit to decide where to invest effort in correcting issues, where to just document them, and where to postpone activities. The poster presents preliminary considerations at the ETH Data Archive of ETH-Bibliothek, the main library of ETH Zurich, on how to address file format identification and validation issues. The underlying issues are mostly independent of the specific tools and systems employed.

KEYWORDS

File format identification; Format validation; Technical metadata extraction; Ingest; Decision making; Preservation planning.

1. INTRODUCTION

To facilitate preservation actions in the future, digital archives rely on comprehensive technical information on file formats being available. Therefore, they try to derive as much information on the characteristics of digital objects as possible already upon or even before ingest. While the processes of format identification, validation or metadata extraction are understood in principle, a number of issues occur in everyday practice. They require an assessment of the specific case followed by a decision on how to proceed without compromising preservation options. Obviously, the broader the spectrum of file formats to be archived and the larger the number of files, the more are scalable efforts required.

One challenge is to understand what kind of issues can be encountered with different types of data. In addition, the tools in use might issue warnings which can also be related to their internal logic. An additional layer is metadata extraction which is also format related, but generally has less immediate effects than identification or validation issues. The practical implications of these issues differ between use cases, customers, types of material, and formats.

2. ETH DATA ARCHIVE

The ETH Data Archive is the institutional data archive of ETH Zurich, a research intensive technical university. We operate the application Rosetta [Ex Libris 2016] as digital preservation system, integrating DROID [The National Archives 2016a] (relying on PRONOM [The National Archives 2016b]) for file format identification and JHOVE [Open Preservation Foundation 2015] for format validation and metadata extraction.

Ingests to the ETH Data Archive comprise research data, administrative records and bequests to the University Archives, and born digital as well as digitized content from the library's online platforms and its digitization center. For research data alone, a broad range of use cases apply, from safeguarding data for a limited period of time (ten years at minimum) to publishing and preserving data in the long term. Several ingest workflows are available to cater for different requirements.

Handling all categories of this varied landscape of use cases adequately is a challenge in many respects. For handling format identification and validation issues, drawing criteria from those use cases' characteristics helps in gaining a better understanding of what actually matters most in each case. Preliminary results are presented in this poster.

3. ISSUES TO BE DISCUSSED

3.1 Format Identification

Ideally, format identification should yield reliable and unambiguous information on the format of a given file. In practice, a number of problems render the process much less straightforward. When it comes to large collections of heterogeneous files in a range of formats, which each may be subject to identification challenges, any effort on the individual files does not scale well. This is a situation we encounter with deposits of research data in particular, but also with bequests of mixed materials to our University Archives. As a result, more or less unsatisfactory decisions need to be taken to keep the volume of data manageable while not rendering potential identification or preservation measures in the future impossible.

3.1.1 Criteria

Example criteria to consider:

- 'Usability': can the file currently be used in the expected way with standard software?
- Tool errors: is an error known to be tool-related?
- Understanding: is the error actually understood?
- Seriousness: is an error concerning the significant properties of the format in question?
- Correctability: is there a straightforward or otherwise documented solution to the error?
- Risk of correcting: what risks are associated with correcting the error?
- Effort: what effort is required to correct the error in all files concerned?
- Authenticity: are there cases where a file's authenticity is more relevant than proper format identification?
- Provenance: is the data producer still available and willing to collaborate in the resolution of preservation issues at least with respect to future submissions?
- Intended preservation level: if bitstream preservation only is expected, the investment into resolving format identification issues might not be justified.
- Intended retention period: if data only needs to be retained for a maximum of ten years, incomplete file format identification might be acceptable.

Obviously, none of these criteria can easily be quantified or translated into simple rules. Even more unfortunately, some of these criteria can actually drive in opposite directions for the same set of files. Therefore, additional questions have evolved:

- Can we continue to handle format identification during ingest into the actual digital archive or will we need to perform it as a pre-ingest activity?
- In the latter case, how would we document in the digital archive measures which are taken prior to ingest to rectify identified problems?
- Under which conditions may we have to admit files with identification fmt/unknown into the archive?
- Should we envisage regular reruns of format identification? If so, how can they be done efficiently and effectively?
- Do we need local format definitions or can we exclusively rely on registries such as PRONOM [The National Archives 2016b] and add information there?
- Is the 'zero applications' risk addressed in any way?

As an indication of the practical and solution independent implications of these issues see e.g. [Mitcham 2015].

3.2 Format Validation and Characterization

File format validation and characterization through metadata extraction are related from a technical point of view. However, the implications of problems in either field can be quite different.

3.2.1 Format Validation

Format validation can fail when file properties are not in accord with its format's specification. However, it is not immediately clear if such deviations prevent current usability of a file or compromise the prospects for a file's long term preservability. If a file can be used readily today, this does not necessarily mean that the file is in itself 'valid enough', either. It rather means that the combination of the file with the application used today is working. This usually requires some generosity in the application's interpretation of the format specification. Obviously, it cannot be assumed that future tools which might have to rely on the documented specification will tolerate such issues. Therefore digital archives need to balance the efforts for making files valid vs. making files pass validation in spite of known issues.

3.2.2 Metadata Extraction

A failure to extract information on significant properties has no immediate consequences, and institutions need to balance the effort in correcting issues. This is even more the case, if embedded metadata or file properties are actually faulty and a correction would involve touching the file itself with a certain risk of unknowingly introducing other changes, too. Based on the criteria listed for format identification, we act therefore even more cautiously when it comes to fixing metadata extraction issues which require a manipulation of embedded metadata or other file properties.

4. ACKNOWLEDGMENTS

We thank all members of the Rosetta Format Library Working Group and in particular the colleagues from the National Library of New Zealand for their ongoing responsiveness and support.

5. REFERENCES

[1] Ex Libris. 2016. Rosetta. (2016). Retrieved July 4, 2016 from http://knowledge.exlibrisgroup.com/Rosetta

[2] Mitcham, Jenny. 2015. File identification ...let's talk about the workflows. (2015). Retrieved July 4, 2016 from http://digital-archiving.blogspot.ch/2015/11/file-identification-lets-talk-about.html

[3] Open Preservation Foundation. 2015. JHOVE - Open source file format identification , validation & characterisation. (2015). Retrieved July 4, 2016 from http://jhove.openpreservation.org/

[4] The National Archives. 2016a. Download DROID: file format identification tool. (2016). Retrieved July 4, 2016 from http://www.nationalarchives.gov.uk/information-management/projects-and-work/droid.htm

[5] The National Archives. 2016b. PRONOM - The Technical Registry. (2016). Retrieved July 4, 2016 from http://apps.nationalarchives.gov.uk/PRONOM/Default.aspx

Web Archiving Environmental Scan

Gail Truman
Truman Technologies, LLC
4096 Piedmont Avenue, Ste. 217
Oakland, CA 94611 USA
+1-510-502-6497
gail@trumantechnologies.com

Andrea Goethals
Harvard Library
90 Mt. Auburn St.
Cambridge, MA 01970 USA
+1-617-495-3724
andrea_goethals@harvard.edu

ABSTRACT

This poster session summarizes the output of a comprehensive Web archiving environmental scan conducted between August and December 2015, with a focus on the preservation-related findings. The scan was commissioned by Harvard Library and made possible by the generous support of the Arcadia Fund.

Keywords

Web Archiving, Preservation, Best Practices

1. INTRODUCTION

Websites are an integral part of contemporary publication and dissemination of information, and as more and more primary source material is published exclusively to the Web, the capture and *preservation* of this ever-growing and ever-changing, dynamic content has become a necessity to support researcher access and institutional needs. Today's research libraries and archives recognize Website archiving ("Web archiving") as an essential component of their collecting practices, and various programs to archive portions of the Web have been developed around the world, within various types and sizes of institutions, including national archives and libraries, government agencies, corporations, non-profits, museums, cultural heritage and academic institutions.

To meet Website acquisition goals, many institutions rely on the expertise of external Web archiving services; others, with in-house staff, have developed their own Web archiving services. Regardless of the approach, the rate at which textual, visual, and audio information is being produced and shared via the Web, combined with the complexity and specialized skills and infrastructure needed for Web archiving processes today – from capture through quality assurance, description, and eventual discovery, to access and analysis by researchers – poses significant resource and technical challenges for all concerned.

Harvard Library sponsored an environmental scan [1] to explore and document current Web archiving programs (and institutions desiring a similar capacity) to identify common concerns, needs, and expectations in the collection and provision of Web archives to users; the provision and maintenance of Web archiving infrastructure and services; and the use of Web archives by researchers. The ultimate goal of the survey was to identify opportunities for future collaborative exploration

This environmental scan is not the first investigation into these areas. Other surveys over recent years have provided valuable information about the landscape of Web archiving activities, such as:

- The National Digital Stewardship Alliance (NDSA)'s Web Archiving in the United States. A 2013 Survey [2]
- NDSA Web Archiving Survey Report, 2012 [3]
- North Carolina State University (NCSU) social media scan, 2015 [4]
- A Survey on Web Archiving Initiatives, Portugal, 2011 [5]
- Use of the New Zealand Web Archive [6]
- Researcher Engagement with Web Archives, 2010 (Dougherty, M) [7]

While there may be overlapping areas covered within these reports and surveys, each examines a particular subtopic or geographical region in relation to Web archiving practices. The NDSA surveys are focused on the USA; the NCSU scan is focused on other areas of social media (such as Twitter) and does not include use cases or details about individual institutions; the Portuguese study examined 42 global Web archiving programs reporting only on the staffing and size (size in terabytes) of each institution's collections; and the Dougherty/JISC study focuses solely on the uses and needs of individual researchers. Other more narrowly focused surveys, such as the IIPC working group surveys, address targeted informational needs.

2. THE SCAN

Through engagement with 23 institutions with Web archiving programs, two service providers and four Web archive researchers, along with independent research, Harvard Library's environmental scan reports on researcher use of – and impediments to working with – Web archives. The collective size of these Web archiving collections is approximately 3.3 PB, with the smallest collection size under one TB and the largest close to 800 TB. The longest-running programs are over 15 years old; the youngest started in 2015. The poster includes the general findings of the scan but emphasizes the findings that are related to preservation.

3. GENERAL FINDINGS

The environmental scan uncovered 22 opportunities for future research and development. At a high level these opportunities fall under four themes: (1) increase communication and collaboration, (2) focus on "smart" technical development, (3) focus on training and skills development, and (4) build local capacity.

4. PRESERVATION FINDINGS

The environmental scan revealed many challenges preserving Web archives - some of them are organizational and some of them are technical. The end result, as one participant put it, is that "Web preservation is at a very immature stage".

The main organizational challenges were knowing whether or not the organization needed to take local responsibility for preservation, being able to trust other organizations to provide preservation services for Web content, lack of funding to pay for the infrastructure demanded by Web archiving, and lack of dedicated staffing with clear roles and responsibilities. Figure 1 shows that more than half of the scan participants report having no dedicated full-time staff for their Web archiving activities.

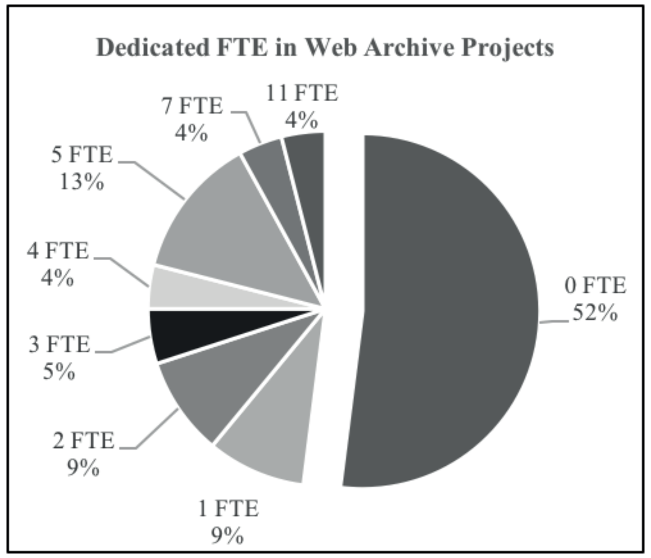


Figure 1. More than half of participants report having no dedicated full-time staff for their Web archive projects.

The technical preservation challenges were largely related to the scale of the Web content being collected and the diversity of the formats captured. Specifically, the main technical preservation challenges were the lack of tools for preparing captured Web content for preservation (see Figure 2); the challenges transferring and storing the large ARC/WARC files; the difficulty capturing certain formats in the first place, particularly social media; the challenges QAing the captures; and the increasing challenges in playing back the Web archives, especially as browsers evolve.

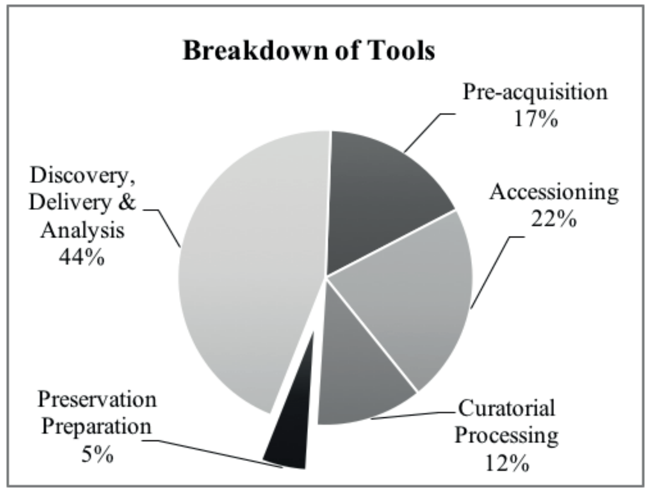


Figure 2. The distribution of tools according to the Web archiving life cycle function.

There were a great number of issues raised relative to the ARC/WARC formats themselves. These difficulties ranged from the complexities of de-duplication, to the difficulty of characterizing the files they wrap, to the difficulties of having to use specialized custom-built tools to process them; and to the problems trying to integrate Web archives with other preserved content.

Art-related Websites frequently break when being archived due to their high levels of dynamic content and interactivity. Preserving that interactivity is currently not possible – and highly desired.

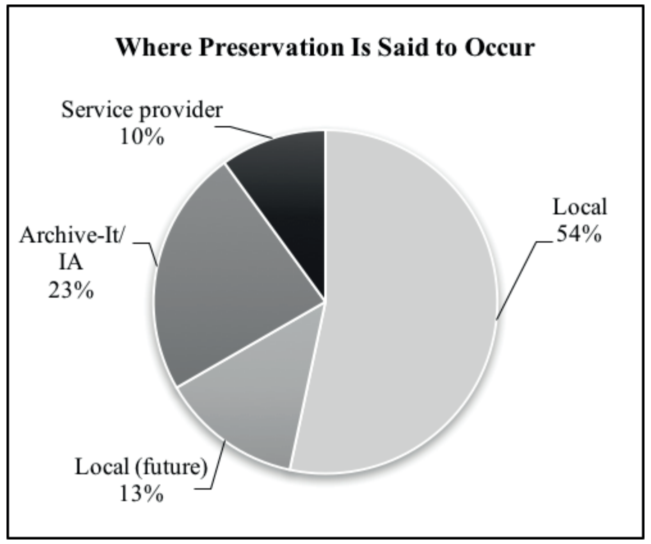


Figure 3. Location of each Web archiving program's preservation copies (now and planned).

5. REFERENCES

[1] Truman, Gail. 2016. *Web Archiving Environmental Scan*. Harvard Library Report. http://nrs.harvard.edu/urn-3:HUL.InstRepos:25658314

[2] National Digital Stewardship Alliance. 2013. *Web Archiving in the United States: A 2013 Survey*. An NDSA Report. http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_USWebArchivingSurvey_2013.pdf

[3] National Digital Stewardship Alliance. 2012. *Web Archiving Survey Report*. An NDSA Report. http://www.digitalpreservation.gov/ndsa/working_groups/documents/ndsa_web_archiving_survey_report_2012.pdf

[4] NCSU Libraries. 2015. Environmental Scan. https://www.lib.ncsu.edu/social-media-archives-toolkit/environment

[5] Gomes, Daniel, Miranda, Joao, and Costa, Miguel. 2011, A Survey on Web Archiving Initiatives. http://sobre.arquivo.pt/about-the-archive/publications-1/documents/a-survey-on-web-archiving-initiatives

[6] National Library of New Zealand. 2015. Use of the NZ Web Archive. http://natlib.govt.nz/librarians/reports-and-research/use-of-the-nz-web-archive

[7] Dougherty, M., Meyer, E.T., Madsen, C., van den Heuvel, C., Thomas, A., Wyatt, S. 2010. Researcher Engagement with Web Archives: State of the Art. London: JISC.

PANELS //

OAIS for all of us

William Kilbride
Digital Preservation Coalition
11 University Gardens
Glasgow G 12 8QQ
+44 (0) 141 330 4522
William@dpconline.org

Sabine Schrimpf
Nestor
Deutsche Nationalbibliothek
Adickesallee 1
D-60322 Frankfurt am Main
+ 49 69 152 517 61
s.schrimpf@dnb.de

Marcel Ras
National Coalition on Digital
Preservation
Prins Willem-Alexanderhof 5
2509 LK Den Haag
+31 6 147 776 71
Marcel.ras@ncdd.nl

Barbara Sierman
KB National Library of the Netherlands
Prins Willem-Alexanderhof 5
2509 LK Den Haag
+31 70 314 01 09
Barbara.Sierman@KB.nl

ABSTRACT

In this panel we will discuss various aspects of the ISO 17421:2012 OAIS standard.

CCS Concepts

Information systems→Standards

Keywords

OAIS; digital preservation; standards; community building.

1. INTRODUCTION

Without exaggeration we could state that the OAIS standard has been most influential in the development of digital preservation within the last decades. With its definition of terms, concepts and models, the OAIS reference model has shaped the way we talk and think about digital archiving in general and about the systems we use for this task. It has become the lingua franca of digital preservationists and one of the most important and most influential documents in the field of digital archiving. Besides that, a range of standards have emerged around and related to OAIS including PREMIS (for preservation metadata), Data Seal of Approval, DIN 31644 and ISO16363 (for certification) and PAIMAS (for exchange between Producers and Archives).

2. UPTAKE AND REVIEW

Since OAIS was initially proposed the digital preservation community has grown tremendously in absolute numbers and in diversity. OAIS adoption has expanded far beyond the space data community (as its original publisher of the standard) to include cultural heritage, research data centres, commerce, industry and government. The needs of users, the technologies in question and the scales of data have also expanded beyond recognition.

The upcoming ISO review of the OAIS standard in 2017 offers a chance for a cooperative review process in which all domains should be involved. Those who depend on the standard have an opportunity to modernise it: those who have found it unwieldy or incompatible with their needs have a chance to report their concerns. It also creates an opportunity for further community building around digital preservation standards, and create even more awareness and understanding. In order to support community building in digital preservation, a community forum via a wiki has been set up. This wiki could be used by professionals to provide feedback on the standard and discuss topics to be fed into the review process. It will provide us with a common view on the state of digital curation and preservation and provide the basis for contributions to the OAIS review.

The outcome from this activity is not simply a wiki nor the ability of stakeholders to generate informed recommendations for change. By providing a shared open platform for the community that gathers around the OAIS, the initiators of the wiki aim to ensure on-going dialogue about our standards and their implementation in the future.

In response to the OAIS community forum, several national initiatives have started to support and contribute to the initiative. DPC in the UK, nestor in Germany and the NCDD in The Netherlands all have set up small working groups to address areas that might worth looking at in the upcoming revision process and will feed them into the forum.

The panel session for iPRES 2016 will give an overview of these national discussions and contributions and will give a floor to continue the discussion with the preservation community based on the themes articulated by the working groups. As the OAIS standard is basis of our daily work, and all of us have opinions about it, the revision of the standard brings our community a perfect opportunity to give input to the standard. Because only if the interested, and qualified, parties voice their issues and concerns, the revision process will yield an appropriate result. We believe that iPRES is the best place we could think of for discussion, community building and adding the issues and concerns of our community to the revision process.

The panel session will be jointly organized by DPC (William Kilbride and Paul Wheatley), nestor (Sabine Schrimpf) and NCDD (Marcel Ras), together with the KB (Barbara Sierman) and the UK Data Service and UK Data Archive (Hervé L'Hours).

Software Sustainability and Preservation: Implications for Long-term Access to Digital Heritage

Jessica Meyerson
Briscoe Center for American History
University of Texas at Austin
2300 Red River, Austin, TX 78712
j.meyerson@austin.utexas.edu

David Rosenthal
Stanford University
518 Memorial Way
Stanford, CA 94305
dshr@stanford.edu

Euan Cochrane
Yale University
344 Winchester Avenue
New Haven, CT 06520
euan.cochrane@yale.edu

Zach Vowell
California Polytechnic State University
San Luis Obispo, 1 Grand Avenue,
San Luis Obispo, CA 93407
zvowell@calpoly.edu

Natasa Milic-Frayling
UNESCO PERSIST Programme
Computer Science Department
University of Nottingham, UK
psznm@nottingham.ac.uk

ABSTRACT

Digital content and data require software for interpretation, processing, and use. This requirement raises the issue of sustaining software functionality beyond its prime use, when it is fully supported and maintained. Virtualization and emulation are two techniques that can encapsulate software in its functional form; furthermore, emulation has recently gained traction as a viable option for long-term access to digital objects. At the same time, archivists, librarians, and museum curators have begun concerted efforts to preserve software that is essential for accessing the digital heritage. In this context the members of the panel will discuss relevant work that they have been involved in to address the goal of software sustainability and preservation.

KEYWORDS

Software preservation, distributed digital preservation networks, partnerships, digital cultural heritage, fair use, copyright

1. THEME

1.1 Software Sustainability

As demand for a specific software application declines, it is typically not economically viable for the software vendor to continue maintaining it for use within contemporary computing environments. Yet, the software may be needed at the later time to access digital content that requires it or which is most authentically rendered using it.

Virtualization and emulation can hold software static in time from a maintenance perspective, minimizing economic burden to the software vendor, while also enabling the preservation of access to the software in a form that is usable in contemporary computing environments.

David Rosenthal's 2015 report to the Mellon Foundation acts as a watershed moment for the viability of emulation, and simultaneously articulates the challenges faced by emulation practitioners [6]. Among them are technological and legal aspects that the UNESCO PERSIST project attempts to address [10], and equally complex issues of standards and established practices that need to be revisited in view of new technical capabilities provided by emulation tools and frameworks such as the bwFLA project from the University of Freiburg [3], the Olive project from Carnegie Mellon University [5], and the JSMESS work being advocated for and implemented by the Internet Archive [8].

1.1.1 Technical Feasibility

The 2009-2012 KEEP Emulation Framework project [2] provided the first "simple" framework for libraries, archives and museums to use in providing access to content via pre-configured emulated or virtualized computers. Since 2012 the bwFLA Emulation as a Service (EaaS) project has been demonstrating remote access to emulated and virtualized environments, accessible via a simple web browser. In addition to a "generic Application Programming Interface (API)" for a variety of emulation and virtualization tools, the team has implemented emulation and virtualization as part of operational archival and library workflows. Through the use of sophisticated virtual disk management and "derivative environment" management, the bwFLA EaaS framework can support a highly distributed software preservation ecosystems that may be attractive to software IP holders. The bwFLA team has also demonstrated automated large scale migration of content using emulated software to perform migration via a "simple" interface.

Mahadev Satyanarayanan demonstrated the feasibility of installing, searching, and streaming VMs with executable content, making it easy to share and instantiate legacy digital artefacts [4]. Natasa Milic-Frayling [1] demonstrated the hosting of legacy software in a contemporary (commercial) cloud platform. Combined with the scalable format transformation services, built as part of the SCAPE EU project [7], software virtualization provides the full range of cloud capabilities for rendering digital content, from 'authentic' content using original software applications to migrated content using contemporary applications.

1.1.2 Content Preservation Practices

New capabilities are prompting memory institutions to revisit current practices and standards in content preservation. This is essential for both the quality of preservation and the development of a market for supporting services. With the increased volume and computational complexity of digital artefacts, it is expected that combined emulation and migration approaches will become a common practice. Moreover, the increased acknowledgment of the feasibility and scalability of emulation tools and services is beginning to shift how cultural heritage institutions approach their preservation strategies. In particular, the economics of emulating content when necessary now presents an attractive alternative to the policies of migrating all digital content over time. Thus, it is important to reflect on the implications of emulation viability for memory institutions, including preservation standards and the development and support of new services.

1.2 Software Preservation

Memory institutions and software vendors possess software collections that present valuable digital heritage and require due care. Furthermore, a growing number of digital objects are software-dependent, i.e., software is essential for their faithful rendering and use. Through research and informal discussions with various stakeholders, the Software Preservation Network (SPN) project [9] has demonstrated and verified that information professionals are confronting such software dependence now.

Both researchers and practitioners have engaged in projects and initiatives to create essential resources and establish effective practices in software preservation, from metadata frameworks to technical capabilities required to create software images from obsolete storage media. In addition, Yale University, SPN, and the Society of American Archivists' CAD/BIM Task Force have pursued relationships with software rights holders in order to resolve legal impediments to preservation practices. Generally, the preservation community continues to evolve their practices and strive for more comprehensive and complete technical registries to support and coordinate software preservation efforts.

2. PROGRAM

The panel will be structured to include a brief introduction of the overall topic, followed by panelist reports on software preservation initiatives over the past 2 years. The reports will be followed by a discussion on the topic with the audience, moderated by Maureen Pennock, Head of Digital Preservation at the British Library. Both the reports and discussion will entice the audience and panelists to reflect and take part in the discussion of several aspects of software preservation and legacy software services:

Community coordination. How to leverage ongoing developments towards a coordinated effort to collect and preserve software essential to access our digital heritage?

Legacy software licenses. How to approach legal issues related to commercial and orphan legacy software?

Economic sustainability. What evidence is required for market viability? Can cultural heritage institutions make a business case to rights holders for preserving software?

Technology infrastructure. Implementation, management, and access to legacy software services.

Standards and best practices. Development of guidelines for cultural heritage institutions that need to re-use software.

3. PRESENTERS AND PANELISTS

Zach Vowell is the Digital Archivist at the Robert E. Kennedy Library, California Polytechnic State University, San Luis Obispo, and co-primary investigator on the Software Preservation Network (SPN) project. Vowell will discuss the input from the research and cultural heritage community, gathered by the SPN team, and the implementation roadmaps developed at the SPN forum in August 2016.

Euan Cochrane is the Digital Preservation Manager at Yale University Library. He has worked with emulation tools since 1990s and collaborated with the University of Frieburg on digital forensics since 2011. Cochrane is currently working on the citation framework for complex and interactive digital objects, funded by DFG/NEH. He will present on emulation use to establish legal access to CD-ROMs from the Yale Library and preservation of canonical versions of installed software environments.

Jessica Meyerson is the Digital Archivist at the Dolph Briscoe Center for American History, University of Texas at Austin, and co-primary investigator on the SPN project. Meyerson will address the legal issues confronted by the SPN team, as well as the implementation roadmaps developed at the SPN forum in August 2016.

David Rosenthal is co-founder and Chief Scientist of the LOCKSS Program at the Stanford Libraries. He will discuss report conclusions and activities following its release [6].

Natasa Milic-Frayling is Chair of the Technology and Research Workgroup of the UNESCO PERSIST Programme and Prof. and Chair in Data Science at the University of Nottingham, UK. She will present on the PERSIST Programme and the plans to create a platform for hosting operational installations of legacy software.

4. ACKNOWLEDGMENTS

Outcomes and results from the Software Preservation Network project were funded by the Institute for Museum and Library Sciences (IMLS), grant number LG-73-15-0133-15.

5. REFERENCES

[1] Abe, Y., Geambasu, R., Joshi, K., Lagar-Cavilla, A., and Satyanarayanan, M. "vTube: Efficient Streaming of Virtual Appliances Over Last-Mile Networks". Proc. of the ACM Symposium on Cloud Computing 2013, Santa Clara, CA

[2] "About the Emulation Framework". *KEEP Emulation Framework [EF]*. The KEEP Project, 2011. Web. 13 July 2016. <http://emuframework.sourceforge.net/about.html>

[3] *bwFLA Emulation as a Service*. University of Freiburg, n.d. Web. 13 July 2016. <http://eaas.uni-freiburg.de/>

[4] Milic-Frayling, N. "Sustainable Computation—Foundation for Long term Access to Digital". Workshop on Open Research Challenges In Digital Preservation, iPRES 2013, Lisbon, Portugal

[5] *Olive Archive*, Olive Archive, 2015. Web. 13 July 2016. <https://olivearchive.org>

[6] Rosenthal, D. S.H. "Emulation and Virtualization as Preservation Strategies". Report prepared for The Andrew W. Mellon Foundation, October 2015, New York, NY

[7] *SCAPE: Scalable Preservation Environments*. The SCAPE Project and the Austrian Institute of Technology, 2014. Web. 13 July 2016. <http://www.scape-project.eu>

[8] Scott, Jason. "Still Life, With Emulator: The JSMESS FAQ". *Internet Archive Blogs*. The Internet Archive, 31 Dec. 2013. Web. 13 July 2016. <https://blog.archive.org/2013/12/31/still-life-with-emulator-the-jsmess-faq/>

[9] *Software Preservation Network*. The Software Preservation Network Project, 2016. Web. 13 July 2016. <http:/www.softwarepreservationnetwork.org>

[10] van Gorsel, M., Leenaars, M., Milic-Frayling, N., and Palm, J. "Evaluation and Strategies of Digital Preservation & UNESCO's Role in Facing the Technical Challenges". 2nd Ann. Conf. of the ICA, UNESCO-PERSIST, 2014, Girona, Spain

Challenges and benefits of a collaboration of the Collaborators

How can we learn and benefit from experiences and set next steps in international collaboration

William Kilbride
Digital Preservation Coalition
11 University Gardens
Glasgow G12 8QQ
+44 (0) 141 330 4522
William@dpconline.org

Sabine Schrimpf
Nestor
Deutsche Nationalbibliothek
Adickesallee 1
D-60322 Frankfurt am Main
+ 49 69 152 517 61
s.schrimpf@dnb.de

Joachim Jung
Open Preservation Foundation
66 Lincoln's Inn Fields,
London, WC2A 3LH,
United Kingdom
joachim@openpreservation.org

Marcel Ras
National Coalition on Digital
Preservation
Prins Willem-Alexanderhof 5
2509 LK Den Haag
+31 6 147 776 71
marcel.ras@ncdd.nl

ABSTRACT

From the very beginning, the importance of collaboration in digital preservation was emphasized by many professionals in the field. Facing the rapid technological developments, the growing amount of digital material and the growing complexity of digital objects, it seems clear that no one institution can do digital preservation on its own. Digital preservation related tasks and responsibilities concern multiple spheres of competence, and over the last two decades, a network of relationships has grown between the stakeholders involved.

Collaborations, from the start, were often driven by research and development issues and framed within large-scale national and international projects delivering usable results for organizations.

In 2002, the DPC was founded as a "collaborative effort to get digital preservation on the agenda of key decision-makers and funders" (Beagrie, 2001). Similar intentions led to the foundation of nestor (2003), NCDD (2007), and NDSA (2010). OPF (2010) and PrestoCenter (2012) were set up as international competence centers. Overall, these organizations serve as platforms for training, knowledge exchange and the study of specific preservation related issues.

These organizations are mainly established to be an advocate and catalyst for cross-domain collaboration in digital preservation. Being networks of competence for digital preservation in which libraries, archives, museums and experts from other domains work together to ensure the long-term preservation and accessibility of digital sources. As is stated perfectly in DPC's mission:

We enable our members to deliver resilient long-term access to digital content and services, helping them to derive enduring value from digital collections and raising awareness of the attendant strategic, cultural and technological challenges they face. We achieve our aims through advocacy, workforce development, capacity-building and partnership.

A logical next step, at least on a national level, is to establish a collaborative infrastructure in which to preserve all relevant digital data from the public sector. To achieve this, we need more than just collaborative storage facilities. Crucially, knowledge and manpower are required to ensure proper management of these facilities. Furthermore, solid agreements must be reached about our various responsibilities: which tasks ought to be performed by the institutions themselves, and which can be carried out in collaboration with others.

The level of maturity in digital preservation, where now some basic principles are established, will also contribute to this development. Collaboration is crucial to achieve this, on a national and cross-domain level, but also on an international level. Therefore the organizations focusing on collaboration are increasingly seeking collaboration with each other. NCDD and DPC are already working together based on a friendship agreement, as do DPC and nestor and DPC and NDSA. OPF has a Memorandum of Understanding with DPC and in the near future with nestor and NCDD.

As the collaboration between the collaborative organizations is starting to come about, this seems the right time to discuss the opportunities we have in joining these forces and being more effective in our work.

In this panel we want to report about the challenges and benefits that we experience in our collaborative efforts. We will encourage the audience to add their experiences with initiatives like ours. We will furthermore ask input from the audience, what issues should we prioritize? From which activities does the community profit the most? Should we focus on practical solutions or on policy making (or on both)? NCDD, DPC, NDSA, Nestor, OPF and PrestoCentre all have their own communities, programs, ideas and communications which form the basis for their activities. These could be expanded to a "collaboration of the collaborators". The panel will present some ideas for such a collaboration and it will invite the preservation community to bring in ideas and issues for next steps in international collaboration.

The proposed session will be jointly organized by DPC (William Kilbride), nestor (Sabine Schrimpf), OPF (Joachim Jung), NDSA (Micah Altman and Nancy McGovern) and NCDD (Marcel Ras).

Beagrie, 2001: Neil Beagrie: Towards a Digital Preservation Coalition in the UK. Ariadne, Issue 27, March 2001. http://www.ariadne.ac.uk/issue27/digital-preservation/

Keywords

Digital Preservation, community building, collaboration

WORKSHOPS & TUTORIALS //

What is Preservation Storage?

Andrea Goethals
Harvard Library
90 Mt. Auburn St.
Cambridge, MA 01970 USA
+1-617-495-3724
andrea_goethals@harvard.edu

Steve Knight
National Library of New Zealand Te Puna Mātauranga o Aotearoa
Cnr Molesworth & Aitken St
Wellington, New Zealand
+64 21 490 503
steve.knight@dia.govt.nz

Jane Mandelbaum
Library of Congress
101 Independence Ave
Washington, DC 20540 USA
+1-202-707-4894
jman@loc.gov

Nancy McGovern
MIT Libraries
77 Massachusetts Ave (14E-210)
Cambridge, MA 02139 USA
+1-617-253-5664
nancymcg@mit.edu

Gail Truman
Truman Technologies
4096 Piedmont Avenue, Suite 217
Oakland, CA 94611 USA
+1-510-502-6497
gail@trumantechnologies.com

Kate Zwaard
Library of Congress
101 Independence Ave
Washington, DC 20540 USA
+1-202-707-5242
kzwa@loc.gov

Sibyl Schaefer
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093 USA
+1-858-246-0741
sschaefer@ucsd.edu

ABSTRACT

Storage is arguably the most fundamental building block of any technical infrastructure designed to preserve digital content. All institutions responsible for preserving digital content must use storage systems or services, whether they are maintaining their own, sharing infrastructure with other institutions, or relying on a third party to provide this service. There are many decisions to be made as to what constitutes a storage infrastructure good enough to protect the integrity and viability of the content, at a cost affordable to the institution. To date each institution has had to independently make these decisions without guidance from the digital preservation community. This workshop will explore whether or not it is possible for the attendees to write a brief "principles of preservation storage" as a starting point for a larger discussion within the community after the conference.

Keywords

Digital Preservation Storage; Preservation Best Practices

1. PRESERVATION STORAGE

Storage is arguably the most fundamental building block of any technical infrastructure designed to preserve digital content. All institutions responsible for preserving digital content must make use of storage systems or services, whether they are maintaining their own, sharing infrastructure with other institutions, or relying on a third party to provide this service. All storage is not equivalent. There are many decisions to be made as to what constitutes a storage infrastructure good enough to protect the integrity and viability of the content, at a cost affordable to the institution. To date each institution has had to independently make these decisions without the help of guidance from the digital preservation community. This workshop will explore whether or not it is possible for the attendees to write a brief "principles of preservation storage" as a starting point for a larger discussion within the community after the conference.

2. WORKSHOP

Prior to the workshop, the authors of this proposal will use mailing lists and social media to solicit requirements documents, RFPs, or other documents that contain institutional requirements for preservation storage. The authors will aggregate the requirements found in the documents, normalizing and de-duplicating them into a single list of unique requirements that could potentially be applicable to any institution preserving content. This list will be transformed into a short Qualtrics survey asking respondents to categorize each requirement as either very important, somewhat important, somewhat unimportant, or not important. Those who sign up for the workshop when they register for the conference will be asked to fill out this survey. The authors will analyze the results to prepare for the workshop discussion. Some of the key points of the discussion will be:

- Which requirements do most agree are very important?
- Which requirements do most agree are not important?
- Where is there a lot of disagreement, e.g. requirements that many thought were very important and many thought were not important? What are the reasons on both sides?
- For which requirements does the group have vague opinions, e.g. most thought it was somewhat important or somewhat unimportant?
- What additional conditions influence requirements? For example, the extent and nature of content types, capacity/shear volume or size of individual "blobs"/file sizes, country or geographic law or industry regulations?

After the workshop discussion the authors will produce a white paper describing the results of the survey and summarizing the workshop discussion.

Personal Digital Archiving: How Can the Cultural Heritage Community Help Individuals Curate Their Own Materials?

Maurizio Lunghi
Arca Memorie non profit association
Florence, Italy, 50100
www.arcamemorie.net
+39 3351396371
mauriziolunghi66@gmail.com

Helen Tibbo
School of Information and Library Science, University of North Carolina at Chapel Hill, NC 27599-3360
+1 919 962 8063
Tibbo@ils.unc.edu

Natascha Schumann
GESIS – Leibniz Institute for the Social Sciences
50667 Cologne, Germany
+49-(0)221-47694-423
natascha.schumann@gesis.org

WORKSHOP DURATION

One-half day. This will include presentations from 6 (or more) speakers plus time for questions and audience discussion.

ABSTRACT

Personal Digital Archiving (PDA) is a relatively recent topic in the landscape of digital preservation and new questions/challenges arise as devices, tools, and apps to capture and share information seem to appear every day. Individuals and community organizations can be overwhelmed with photos, email messages, texts, letters, and a wide array of other materials. This workshop seeks to discuss ways in which cultural heritage institutions such as libraries, archives, and museums, along with university researchers, and software and systems developers in this domain, can help individuals and groups come to grips with their digital collections and preserve content that is important to their lives, organizations, communities, and heritage in trustworthy, long-term ways. Very few initiatives are on-going worldwide on this topic, one of the most relevant, the PDA conference in the US, is presented to bring the discussion to Europe.

Cultural heritage organizations that are building trustworthy digital repositories along with tool builders and software/system developers have much to offer in this arena but roles and responsibilities as well as incentives and resources must be established. At the most fundamental stage, awareness needs to be raised among all stakeholders and guidelines to help individuals and organizations need to be developed, maintained, and disseminated. Of prime concern at this juncture is how public institutions such as libraries and archives can help community organizations and individuals?

Keywords

Personal Digital Archiving; Cultural Heritage; Selection Criteria; Preservation Planning.

1. WORKSHOP ORGANISERS

- Maurizio Lunghi, Arca Memorie non-profit association
- Helen Tibbo, School of Information and Library Science, University of North Carolina at Chapel Hill
- Cal Lee, School of Information and Library Science, University of North Carolina at Chapel Hill
- Natascha Schumann, GESIS – Leibniz Institute for the Social Sciences
- Achim Osswald, TH Köln / Technology, Arts, Sciences, Institut fuer Informationswissenschaft
- Martin Iordanidis, Gruppe Publikationssysteme, Hochschulbibliothekszentrum des Landes NRW (hbz)

2. WORKSHOP TARGET AUDIENCE

- Community organizations focused on gathering oral histories or other local collections;
- Representatives of cultural heritage institutions who are interested in helping individuals preserve their personal digital contents as well as oral history and other local collections;
- University scholars, researchers, and students working in areas related to Personal Digital Archiving including archival science, information systems, oral history, privacy and other legal area, and other related fields;
- Private companies or associations interested in offering/facilitating technical solutions and infrastructures suitable for a long-term archival services for individuals and community and local collections;
- Information professionals including archivists, librarians, and curators; and
- Those preserving family material, activist groups, and hobbyists.

3. WORKSHOP DESCRIPTION AND OBJECTIVES

Personal Digital Archiving (PDA) is a relatively recent topic in the landscape of digital preservation and new questions/challenges arise as devices, tools, and apps to capture and share information seem to appear every day. Individuals and community organizations can be overwhelmed with photos, email messages, texts, letters, and a wide array of other materials. This workshop seeks to discuss ways in which cultural heritage institutions such as libraries, archives, and museums, along with university researchers, and software and systems developers in this domain, can help individuals and groups come to grips with their digital collections and preserve content that is important to their lives, organizations, communities, and heritage in trustworthy, long-term ways.

This workshop will explore the domain of Personal Digital Archiving by defining potential actors and roles, and discussing key topics such as resources, outreach, privacy, legal issues, and technical options available for individuals and organizations to preserve their digital content. Presently, few institutions around the globe are mounting concerted efforts to help individuals curate their own materials, however, there has been a PDA conference in the US each year since 2010. It is our goal in this workshop to bring this discussion to iPRES2016 and to Europe. Because of the newness of PDA efforts it is important to be inclusive and bring as many voices to the conversation as possible.

For many artists, writers, politicians and other well-known figures, archives and other public institutions have preserved their personal papers and other artifacts. However, these institutions have saved very little material from average citizens. While families may have had success preserving their own paper-based content from generation to generation, digital media present an entirely new and often confounding range of problems for those who wish to preserve items such as wedding videos, Facebook pages, or even email messages. Few individuals or community organizations today understand the fragility of digital media or have any knowledge as to how to preserve files or migrate content to future formats.

Many open questions must be addressed before most small organizations and individuals will have secure preservation services for their content. Cultural heritage organizations that are building trustworthy digital repositories along with tool builders and software/system developers have much to offer in this arena but roles and responsibilities as well as incentives and resources must be established. At the most fundamental stage, awareness needs to be raised among all stakeholders and guidelines to help individuals and organizations need to be developed, maintained, and disseminated. Of prime concern at this juncture is how public institutions such as libraries and archives can help community organizations and individuals? Recent discussions concerning a public-private collaboration among cultural heritage institutions and systems builders to deliver a sort of “subsidiary service” for preserving individuals’ digital memories is one example of possible paths toward realizing PDA for the public.

One of these new services is proposed by the Arca Memorie, a non-profit association www.arcamemorie.net designed to help individuals preserve their digital memories, where the members can use a repository service with a guarantee of long-term preservation of their digital content. The key points of the service are the ownership of the digital objects, the aspects related to long term preservation including digital formats, contextual metadata, persistent identifiers, authenticity, and provenance. A field trial has been launched in April 2016 and is still open.

In order to customize the workshop for participants, the organisers will collect information through an online questionnaire before the event concerning participants’ general understanding of personal archiving, what they see as their roles in PDA, what they (or their users) want to preserve and why, the possible goals of such preservation activity, user expectations or requirements, the legal or economic constraints, and so on. The collected data will be used to stimulate the discussion at the workshop and the participants will be invited to contribute forward. The workshop is expected to be a first step of a long path toward clarification of the PDA topics and possible solutions.

Quality Standards for Preserving Digital Cultural Heritage

Börje Justrell
Riksarkivet
Fyrverkarbacken 13 -17, Stockholm, Sweden
+46104767187
borje.justrell@riksarkivet.se

Antonella Fresa
Promoter Srl
Via Boccioni, 2, Peccioli (PI), Italy
+390587466881
fresa@promoter.it

Claudio Prandoni
Promoter Srl
Via Boccioni, 2, Peccioli (PI), Italy
+390587466881
prandoni@promoter.it

Peter Pharow
Fraunhofer IDMT
Ehrenbergstrasse 31, 98693 Ilmenau, Germany
+493677467363
phw@idmt.fraunhofer.de

Magnus Geber
Riksarkivet
Fyrverkarbacken 13 -17, Stockholm, Sweden
+46702561317
magnus.geber@riksarkivet.se

Erwin Verbruggen
Nederlands Instituut voor Beeld en Geluid
Media Parkboulevard 1, 1217 WE Hilversum, The Netherlands
+31624151991
everbruggen@beeldengeluid.nl

ABSTRACT

Memory institutions face increasing volumes of electronic documents and other media content for long term preservation. Data are normally stored in specific formats for documents, images, sound, video, etc., produced by software from different vendors. This software is controlled neither by the institution producing the files, nor by the institution that archives it. This obligates memory institutions to carry out conformance tests before accepting transfers of electronic collections, but again these are beyond the control of the institution and can be unreliable. This poses problems for long-term preservation. Digital content, meant for preservation, passing through an uncontrolled generative process can jeopardise the preservation process. The objective of PREFORMA (PREservation FORMAts for culture information/e-archives) – a Pre Commercial Procurement project co-funded by the European Commission under its FP7-ICT Programme – is to give memory institutions full control of the conformance testing of files created, migrated and ingested into archives. This is achieved through the development of open source tools which enable this process within an iteration that is under the full control of memory institutions. The project aims to establish a sustainable ecosystem involving interested stakeholders from a variety of backgrounds, including researchers, developers and memory institutions. The workshop will present the results of the project, including demonstration of the conformance checkers developed during the prototyping phase. This will inform a discussion with the digital preservation community – open source community, developers, standardization bodies and memory institutions – about the opportunities offered by PREFORMA and the challenges that are still to be addressed.

Keywords

standard file formats; digital cultural heritage; open source; digital preservation; conformance tests.

1. INTRODUCTION

PREFORMA (www.preforma-project.eu) is an EU-funded Pre-Commercial Procurement project working on one of the main challenges faced by memory institutions: the **long-term preservation of digital data**.

The main objective of the pre-commercial-procurement launched by PREFORMA is the development and deployment of three **open source tools**. These are developed for memory institutions, or any organisation with preservation requirements, wishing to check file-format conformance with a specific technical standard. The conformance checkers establish whether a file complies with a specific specification and facilitate memory institutions' acceptance criteria outside of file-format specifications. The software reports in human and machine-readable format which properties deviate from the specification and acceptance criteria, and can perform automated fixes for simple deviations in the file's metadata.

The process of conformity testing guarantees that the digital content to be ingested into the archives conforms to standards and, if necessary, the content can be re-processed for corrections.

The veraPDF consortium, Easy Innova and MediaArea are working on the implementation of the conformance checkers. The software development is carried out in a collaborative environment with memory institutions and subject-matter experts. The tools are being developed in an iterative process including a number of experiments with 'real' data sets from memory institutions and other organisations who have offered to participate in the testing phase, with the aim of demonstrating the software's effectiveness and refining the prototype releases.

The standards[1] covered in the project are:

- PDF/A for electronic documents
- uncompressed TIFF for images
- a combination of MKV, FFv1 and LPCm for AV files.

The first prototypes are available for download from the PREFORMA Open Source Portal at http://www.preforma-project.eu/open-source-portal.html.

Due to the need for sound, long-term sustainable solutions, the overall objective of the PREFORMA project is not "artefact centric", but instead aims to establish a long-term sustainable ecosystem around the developed tools, involving interested stakeholders from a variety of backgrounds.

2. TARGET AUDIENCE

The workshop is a step towards the establishment of this self-sustaining community and is aimed at anyone interested in digital preservation and cultural heritage:

- *Memory institutions* (museums, libraries, archives, etc.) and cultural heritage organisations coordinating or representing memory institutions that are involved in (or planning) digital culture initiatives.
- *Developers* who contribute, or are interested in contributing code to the PREFORMA tools as well as those developing commercial solutions and *enterprises* integrating the reference implementations into production software.
- *Research organisations* providing technical and expertise advice to cultural stakeholders.
- *Standardisation bodies* maintaining the technical specifications for the curation and preservation formats covered in PREFORMA.
- *Funding agencies*, such as Ministries of Culture and national/regional administrations, that own and manage digitisation programmes and may endorse the use of the PREFORMA tools in the digitisation, curation and preservation process.
- *Best practice networks* endorsing the use of open standards for creating and managing digital content.
- *Other projects* in the digital culture, e-Infrastructures and policy arenas that are considering the use of PCP.

The anticipated impact of PREFORMA is the reduction of curation and preservation costs, the improvement of curation and preservation capacity, increased competence in public organisations, including small archives, and enhanced independence from individual vendors.

3. PROGRAMME

The workshop will feature presentations by representatives from the PREFORMA project, live demonstrations of the conformance checkers and a round table discussion which will offer the audience the opportunity to share their views on possible improvements to the tools, possible integrations with other software / systems and what they perceive as the main challenges to be addressed in the future.

Draft Programme:

- 30 mins: presentation of PREFORMA, its main achievements and how to contribute
- 60 mins: working session with demonstrations of the conformance checkers
- 30 mins: break
- 60 mins: round table discussion: integration opportunities and future challenges
- 30 mins: lessons learnt and conclusions.

[1] For further information on the standard file format specifications addressed in PREFORMA visit the project's website at http://www.preforma-project.eu/media-type-and-standards.html

OSS4Pres 2.0: Building Bridges and Filling Gaps

Sam Meister
Educopia Institute
sam@educopia.org

Carl Wilson
Open Preservation Foundation
carl@openpreservation.org

Shira Peltzman
UCLA Library
speltzman@library.ucla.edu

Heidi Dowding
Indiana University
heidowdi@indiana.edu

ABSTRACT

In this paper, we describe the structure and contents for the OSS4Pres 2.0 workshop.

Keywords

open-source software, digital preservation, workflows, advocacy, software requirements.

1. SCOPE

Building on the success of the "Using Open-Source Tools to Fulfill Digital Preservation Requirements" workshop held at IPRES 2015, we propose a "oss4pres 2.0" workshop to further the ideas and themes generated. During the first oss4pres workshop, which was well attended and generated dynamic and engaging discussion amongst participants, digital preservation practitioners and open source software tool developers gathered to discuss the opportunities, challenges, and gaps related to developing open source systems and integrating them into institutional systems and workflows. By engaging in a series of focused activities designed to build upon our findings, at oss4pres 2.0 we seek to both move the conversation forward as well as produce actionable results that will directly benefit digital preservation practitioners.

Increased adoption and implementation of OSS tools within institutional digital preservation workflows has resulted in an increased set of knowledgeable practitioners who are seeking to move beyond simple testing and experimentation of tools. Digital preservation practitioners want to make informed tool selection decisions and participate in the process of developing solutions to better automate and integrate OSS tools. The oss4pres 2.0 workshop will provide a highly interactive forum for an audience of digital preservation practitioners, OSS tool developers, and institutional administrators to engage and collaborate to advance the continued development and implementation of OSS tools within the digital preservation community.

Workshop participants will gain:

Increased knowledge of the opportunities, challenges, and gaps related to the development and implementation of OSS in digital preservation workflows

Increased understanding of OSS implementation strategies and experiences at multiple institutions

Experience collaborating with fellow digital practitioners to develop practical solutions to address current OSS challenges and gaps

2. INTENDED CONTENT

The workshop structure and content will be designed to move quickly from idea generation and discussion to producing tangible outputs and will be composed of the following sessions:

1. Introduction and overview of topics

2. Participants will select from a set of proposed topics and form project groups

3. Each group will work through series of exercises specific to each topic following a basic sequence:

4. Discuss and define problem space for selected topic

5. Develop draft plans and/or solutions documents

6. Revise and finalize draft materials

7. Groups will come back together to report on activities and plan next steps

Potential Topics

Develop requirements for an online community space for sharing workflows, OSS tool integrations, and implementation experiences

Draft a one page advocacy guide for practitioners to utilize to communicate benefits and tradeoffs of OSS to administrative stakeholders

Develop functional requirements and/or features for OSS tools the community would like to see developed (e.g. tools that could be used during 'pre-ingest' stage)

Draft a design guide for FOSS tools aimed at ensuring they're designed to integrate easily with digital preservation institutional systems. Topics might include meeting common operational policy criteria regarding packaging, installation and security, preferred mechanisms for integration, e.g. command line wrapping, REST interfaces, programmatic APIs, or documentation issues.

3. ORGANIZERS

Sam Meister is the Preservation Communities Manager at the Educopia Institute, working with the MetaArchive Cooperative and BitCurator Consortium communities. Previously, he worked as Digital Archivist and Assistant Professor at the University of Montana. Sam holds a Master of Library and Information Science degree from San Jose State University and a B.A. in Visual Arts from the University of California San Diego. Sam is also an Instructor in the Library of Congress Digital Preservation Education and Outreach Program.

Shira Peltzman is the Digital Archivist at UCLA Library where she leads the development of their digital archives program. Previously she has worked with a number of cultural heritage organizations around the world including Martin Scorsese's World Cinema Foundation, the British Film Institute, the Bay Area TV Archive, and Carnegie Hall. Shira was a member of the National Digital Stewardship Program in New York's inaugural cohort and holds a Master's Degree in Moving Image Archiving and Preservation from New York University's Tisch School of the Arts.

Carl Wilson is the Technical Lead at the Open Preservation Foundation. An experienced software engineer with an emphasis on software quality through testing. Carl is an open source enthusiast, both as a user and developer. Carl oversees all of the Open Preservation Foundation's technical activities and is responsible for software quality on the veraPDF project, which is led by the Open Preservation Foundation in partnership with the PDF Association.

Heidi Dowding is the Digital Preservation Librarian at Indiana University, where she is currently focusing on infrastructure development and the creation of the Born Digital Preservation Lab. She is active in the development of national initiatives, including APTrust and Digital Preservation Network. Her previous positions include DiXiT Research Fellow at Huygens ING, Library of Congress National Digital Stewardship Resident at Dumbarton Oaks Research Library and Collection, and Reference and Digital Services Librarian at Nazarbayev University. She holds an MLIS from Wayne State University and BA from Michigan State University.

Innovative practices in Digital Preservation Data Life Cycle Management (DLCM): A Swiss Initiative

Pierre-Yves Burgi & Eliane Blumer
Université de Genève
DiSTIC/DIS
Rue du Général Dufour 24,
CH-1204 Genève
+41 22 379 71 45
eliane.blumer@unige.ch
pierre-yves.burgi@unige.ch

Aude Dieudé
EPFL VPAA SISB
RLC D1 220
Station 20, CH-1016 Lausanne
+41 21 69 34784
aude.dieude@epfl.ch

Ana Sesartić
ETH Zürich
ETH-Bibliothek, Digital Curation
Rämistrasse 101, CH-8092 Zürich
+41 44 632 73 76
ana.sesartic@library.ethz.ch

ABSTRACT

In this workshop, we show how researchers and information professionals are supported throughout the data life-cycle in order to ensure innovative and professional digital preservation in Switzerland.

Keywords

Data life-cycle management; research data management; digital preservation; Swiss initiative; best practices

1. INTRODUCTION

Managing research data throughout lifecycles is a key prerequisite not only for effective data sharing but also for efficient long-term preservation. Timely management of research data greatly facilitates their long-term preservation at the end of the research cycle and might save precious time and resources by avoiding unnecessary duplication of work. Data archiving and preservation are important components of a good data management plan and taking ahead of time decisions on requirements for such a preservation process facilitates data re-use and discovery.

Yet, one main question to explore during this workshop will be the following: How can we best support and effectively reach researchers and information professionals on a national level regarding best innovative practices in digital preservation so as to succeed in this ambitious enterprise?

2. PROJECT OBJECTIVES

One of the primary objectives of the Data Lifecycle Management project (DLCM) is to provide sustainable and tangible solutions to implement research data lifecycle management in Switzerland. In doing so, this project aims at consolidating and further developing collaboration, while in a first step promoting coordination between eight Swiss higher education institutions (EPFL, UNIGE, ETHZ, UZH, UNIBAS, HEG, HES-SO, and SWITCH). Building on existing resources and tools at national and international levels, the DLCM team targets the setting up of the needed services that allow efficient managing of active research data, and ensure publication, long-term reference and preservation of subsets of data selected by researchers. At the term of this project, the DLCM team aspires in the near future to offer and propose its resources and services to other institutions, which either lack experience in this field and/or do not have any infrastructure for managing scientific data.

3. PROJECT CONTENT

What could be the best practices to preserve and manage digital materials on a national level? To summarize, the Swiss project encompasses a range of pilot implementations of DLCM solutions, from hard sciences to digital humanities. Guidelines and data management plan (DMP) checklist and templates, which are essential tools for providing researchers with the incentive to manage and be able to preserve their data, have been customized for Switzerland based on pre-existing national and international policies. These resources are currently available through a recently launched national portal: www.dlcm.ch [1]. Since DLCM of research data touches on many questions, which include, but are not limited, to data organization, file formats, metadata as well as legal and regulatory aspects, important outcomes of this project are the training (in person and online) of the end-users and the offering of best practices consulting in crucial DLCM areas.

4. AIM OF THE WORKSHOP

With this background, this workshop aims at providing the unique opportunity to share the insights, experiences and best practices regarding innovative practices regarding long-term preservation of research data. This Swiss DLCM project, with its experience gained from concrete implementations, and which involves at the various partner institutions research units, IT services, and libraries – will serve as springboard for promoting and animating the discussion and debate across countries and continents. The audience for this workshop will target large communities of researchers nationally and internationally.

5. ACKNOWLEDGMENTS

Our thanks go to the entire DLCM team across institutions led by Pierre-Yves Burgi.

6. TARGET AUDIENCE

This workshop is especially designed for researchers, information professionals, practitioners, librarians, IT specialists, as well as everyone interested in digital preservation's best practices.

7. REFERENCES

[1] Data Life-Cycle Management national portal funded by Swissuniversities: www.dlcm.ch.

Workshop: Symmetrical Web Archiving With Webrecorder

Dragan Espenschied & Ilya Kreymer
Rhizome
235 Bowery
NY 10002, USA
+1-212-219-1288
{firstname.lastname}@rhizome.org

ABSTRACT

This paper describes a **workshop** for the novel, open source web archiving tool *Webrecorder*. Until now, web archiving has mainly been thought to be synonymous with "spidering" or "crawling," meaning that a very basic, simulated version of a web browser travels paths of links and storing what it encounters, based on a certain set of rules.

Webrecorder introduces a new web archiving concept, *symmetrical archiving*, which makes use of actual browsers and actual user behavior to archive the web, as well. The software stack used for accessing or replaying archived material is exactly the same as during the capturing process. This allows for unprecedented fidelity in web archiving, , enabling the preservation of items embedded complex, dynamic web applications, while keeping their whole, interactive context as well as any user specific content.

This new approach to web archiving requires new ways of working within institutions; the proposed workshop serves as an introduction to symmetrical archiving, using Webrecorder's emulation-based browsers, defining object boundaries, and transitioning from or augmenting crawler-based archives.

Keywords

web archiving; symmetrical archiving; emulation; collection management; appraisal; access; open source; institutions

1. THE NEED FOR SYMMETRICAL ARCHIVING

Current web archiving tools are based on the assumption that resources are published within a continuous, two-dimensional system, based on a **location**—the URL—and the **time** the resource was accessed.

The reality of the web has changed, as early as the introduction of the Netscape Navigator browser and Netscape Enterprise Server in 1994: The new server allowed for session-based, personalized content being served to users, client-side scripting in the form of Java applets and Javascript turned browsers into virtual machines that could execute complex behaviors. Although many of those innovations did not follow any rule book or standard, they have effectively made the web culturally and economically relevant, transforming it into the technically and culturally dominant medium it is today. At the same time, current web archiving practices and available tools do not sufficiently acknowledge the web being a complex software environment rather than a document retrieval system, making it impossible to create web archives that reflect for example current practices of cross-embedding, personalized services, web applications etc…

While many big data tools exist to analyze information from the web on a large scale, this is missing the way real users actually consume and create material on the web: the affect is only created with the relationships and contexts staying intact.

Taking on this challenge, *Symmetrical Archiving* assumes that the product of any web archiving activity is highly dependent on the actual activities carried out, and the technological context it is happening in. Resources that are only created in the moment when they're accessed require archivists to be conscious about this activity.

Symmetrical Archiving means that for capture and access the same code is executed. Within the open source web archiving platform *Webrecorder*, looking at a live web site, an archived site, or recording a new site, is the same process: real interactions—carried out manually or via automation—are executed via a real browser on live and/or recorded material, with all data flowing through *Webrecorder* software, with a recording component optionally working.

This has implications for how object boundaries are defined, quality assurance is carried out and collections are structured.

2. EMULATED BROWSERS

A key component of Webrecorder is a growing library of real browsers—from current version of Firefox and Chrome to legacy versions of Internet Explorer, Safari, Netscape and Mosaic—that can be used to record and access web archives. This produces the highest possible web archiving fidelity: during recording, data being requested by for example Java, Adobe Flash, Shockwave or Real Media browser plug-ins can be captured; during playback, the same resources can be re-performed within exactly the same software environment they were recorded in, ensuring long term access to complex web materials.

The emulated browsers are presented to users as a hosted emulation service based on carefully pre-configured software environments that can be brought up in arbitrary number on-demand and are accessible via any modern web browser. No special configuration or complicated security precautions are required. The service is a new emulation framework designed specifically for web browsers, which also powers our previous effort, http://oldweb.today/.

3. TARGET AUDIENCE

The workshop is targeted at archiving professionals from all kinds of memory institutions that either are already engaged in web archiving or are planning to start a web archiving program. No prior knowledge of web archiving or certain kinds of tools are necessary.

Collection managers, curators and institutional infrastructure providers (IT) are welcome to join as well.

4. WORKSHOP PROGRAM

The following topic will be discussed in the workshop:

1. Introduction to symmetrical archiving: In-depth discussion of the concept and its consequences for web archiving practice
2. Introduction to Webrecorder, comparing it with existing web archiving tools
3. Creating a collection, from initial curation to publishing, including choosing the right emulated environment
4. Managing collections
5. Using Webrecorder as a service or deploying it within an institution
6. Advanced users: customizing Webrecorder, emulated environments, using Webrecorder components

The workshop can accommodate up to 18 participants and lasts 90 minutes. At the end of the workshop, users will create their own web archives, which they can choose to keep in Webrecorder hosted service or transfer to a different storage medium as best meets their needs.

5. ABOUT WEBRECORDER

Webrecorder is part of Rhizome's digital preservation program.

Rhizome's digital preservation program supports social memory for internet users and networked cultures through the creation of free and open source software tools that foster decentralized and vernacular archives, while ensuring the growth of and continuing public access to the Rhizome ArtBase, a collection of 2,000+ born-digital artworks.

Rhizome is a non-profit organization that commissions, exhibits, preserves, and creates critical discussion around art engaged with digital culture. It is located in New York and an affiliate to the New Museum.

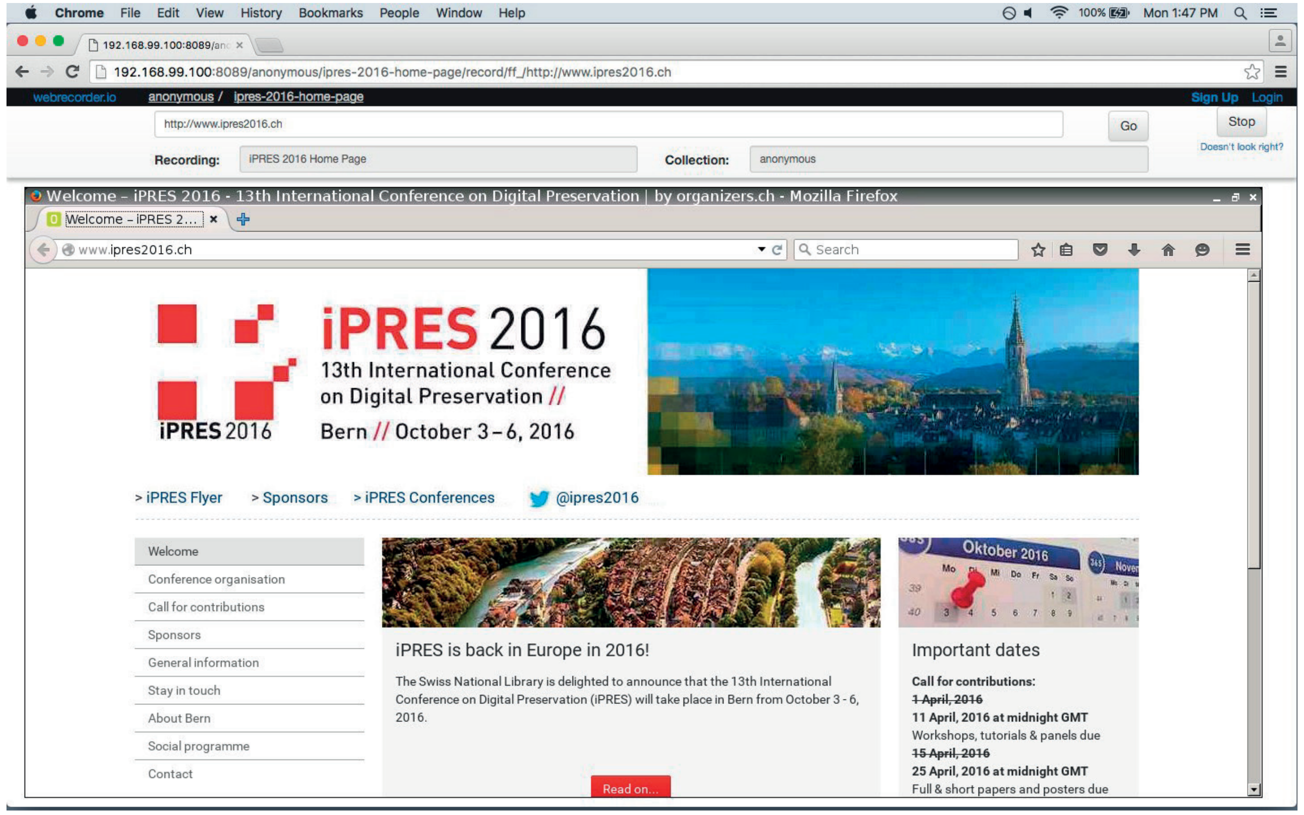
Webrecorder is free, open source software, available at http://github.com/webrecorder/ and as a hosted service at http://webrecorder.io.

Workshop personel:

Ilya Kreymer has previously worked as a software engineer for the Internet Archive and is now the lead developer of Webrecorder at Rhizome.

Dragan Espenschied is the head of Rhizome's Digital Preservation program.

The Webrecorder project receives significant support from the Andrew W. Mellon Foundation.



192.168.99.100:8089/anonymous/ipres-2016-home-page/record/ff_/http://www.ipres2016.ch

http://www.ipres2016.ch

Recording: iPRES 2016 Home Page

Collection: anonymous

Welcome – iPRES 2016 - 13th International Conference on Digital Preservation | by organizers.ch - Mozilla Firefox

www.ipres2016.ch

iPRES is back in Europe in 2016!

The Swiss National Library is delighted to announce that the 13th International Conference on Digital Preservation (iPRES) will take place in Bern from October 3 - 6, 2016.

Important dates

Call for contributions:
~~1 April, 2016~~
11 April, 2016 at midnight GMT
Workshops, tutorials & panels due
~~15 April, 2016~~
25 April, 2016 at midnight GMT
Full & short papers and posters due

Illustration 1: Webrecorder in action, recording the iPRES2016 site in an emulated Firefox browser usable within the user's native Chrome browser.

Persistent Identifiers for Digital Cultural Heritage

Jonathan Clark
International DOI Foundation
Rembrandltlaan 12
1231 AC Loosdrecht
The Netherlands
+31654733572
jonathanmtclark@gmail.com

Remco van Veenendaal & Marcel Ras
National Archives
P.O. box 90520
2509 LM The Hague, Netherlands
+31629451951 / +31703140180
Remco.van.veenendaal@nationaalarchief.nl
Marcel.ras@kb.nl

Maurizio Lunghi
Associazione ARCA MEMORIE
Florence, Italy, 50100
Italy
+393351396371
mauriziolunghi66@gmail.com

Juha Hakala
National Library of Finland
Fabianinkatu 35
00170 Helsinki
Finland
+358 50 382 7678
juha.hakala@helsinki.fi

ABSTRACT

This is an introductory level tutorial. The goals of the tutorial are: to explain what persistent identifiers are and why they are so important; to discuss what the criteria are for trustworthy identifier systems; to introduce a decision tree tool that helps determine which system is the most appropriate for a particular set of needs; and to share case studies that are relevant to the Digital Preservation community.

Keywords

Persistent Identifiers; Handle System; DOI; ARK; URN; URN:NBN; PURL.

1. INTRODUCTION

Over the past years, a growing number of collections belonging to archives, libraries, media, museums, and research institutes have being digitized and made available online. These are exciting times for ALM institutions. On the one hand, they realize that, in the information society, their collections are goldmines, as "data is the new gold" [1]

On the other hand, unfortunately most heritage institutions do not yet meet the basic preconditions for long-term availability of their collections. Apart from the problems of digital preservation the digital objects often have no long lasting fixed reference. URL's and web addresses change. For instance, some digital objects that were referenced in Europeana and other portals can no longer be found. References in scientific articles can have a very short life span, which is damaging for scholarly research, as is shown in the well-known article of Herbert van de Sompel and others [2].

Thus, in this digital world there is a need to unambiguously determine what a resource is and where it can be found in a way that is persistent over time. However, the identifiers themselves are simply strings of numbers and not inherently persistent. What make them persistent are the policies, organization and agreements that sit behind these numbers.

Many different systems have emerged over the years, some more robust than others. What is the difference between them all and how to choose between the various options available? What are the criteria to judge whether a PI system can be trusted or not? [3]

Finally, of course a choice must be made for a PI system that meets the long-term needs of the ALM institution. In 2015, the Dutch Digital Heritage Network (NDE) [4] started a two-year work program to co-ordinate existing initiatives in order to improve the (long-term) accessibility of the Dutch digital heritage for a wide range of users, anytime, anyplace. [5]

The NDE is a partnership established as an initiative of the Ministry of Education, Culture and Science. The members of the NDE are large, national institutions that strive to professionally preserve and manage digital data, e.g. the National Library, The Netherlands Institute for Sound and Vision, the Netherlands Cultural Heritage Agency, the Royal Netherlands Academy of Arts and Sciences, the National Archive of the Netherlands and the DEN Foundation, and a growing number of associations and individuals both within and outside the heritage sector.

Meanwhile, other institutions such as the British Film Institute, the National Diet Library of Japan and the European Union are moving forward with interesting projects to assign persistent identifiers to their archives.

2. TOPICS

The tutorial on persistent identifiers for digital cultural heritage will consist of four parts

1. Introduction: what are persistent identifiers (PI) and why are they important?
   a. Knowing what's what and who's who
   b. The importance of Social Infrastructure
   c. Review of current identifier systems
2. The criteria for reliable & trustworthy PIs
   a. Insights from the APARSEN project [6] [7]
   b. Recommendations of the Research Data Alliance (RDA)[8]

3. The NDE decision tree for PIs
    a. A tool in form of an online questionnaire that guides cultural heritage organizations through the process of selecting a particular type of Persistent Identifier (Handle, DOI or NBN:URN) [9]
    b. Discuss the applicability of the decision tree outside the Netherlands
4. Case studies
    a. NDE from the Netherlands
    b. National Diet Library, Japan
    c. Office of Publication of the European Union

SCOPE

The intention is that this tutorial be "PI-agnostic"; that is, any persistent identifier (Handle [10], DOI [11], ARK [12], URN:NBN [13], URN [14] and PURL [15]) will be discussed.

3. INTENDED AUDIENCE

Anyone interested in learning more about persistent identifiers for digital information. This tutorial has a strong practitioner focus and will be especially interesting for those working with Digital Archives and Digital Collections. This tutorial also acts as an introduction and level set for the Workshop on Smart Persistent Identifiers. Participants may be interested to read some background articles [16] [17] [18] [19]

4. EXPECTED LEARNING OUTCOMES

Participants will leave this tutorial with a clear understanding of what persistent identifiers are and why they are important. They will have an overview of the different identifier systems in use today and have ample opportunity to answer their questions.

5. SPEAKERS

The tutorial speakers will be Jonathan Clark, Managing Agent for the DOI Foundation, Remco van Veenendaal of the Dutch National Archives, Marcel Ras of the National Coalition for Digital Preservation and Maurizio Lunghi of the Associazione ARCA MEMORIE and Juha Hakala of the National Library of Finland.

6. REFERENCES

[1] "Data is the new gold". Is the title of a speech from Neelie Kroes as Vice-President of the European Commission responsible for the Digital Agenda. Opening Remarks on a Conference on Open Data Strategy in Brussels, 12th December 2011. http://europa.eu/rapid/press-release_SPEECH-11-872_en.htm?locale=en

[2] Herbert van de Sompel et al, Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot, Plos One, December 26, 2014. http://dx.doi.org/10.1371/journal.pone.0115253

[3] iPRES 2012, 'Interoperability Framework for Persistent Identifier systems', iPRES 2012, Toronto, urn:nbn:it:frd-9491 https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf; LIBER 43 Annual conference, "Implementing an Interoperability Framework for Persistent Identifier systems" in Riga, Latvia, on 2-5 July 2014, ISBN 978-9984-850-23-8

[4] The Digital Heritage Network in the Netherlands http://www.ncdd.nl/en/ncdd-projects/digital-heritage-network/

[5] Persistent Identifier project information / National Coalition for Digital Preservation http://www.ncdd.nl/en/ncdd-projects/digital-heritage-network/project-persistent-identifiers/

[6] Alliance for Permanent Access to the Records of Science Network http://www.alliancepermanentaccess.org

[7] APARSEN project, Deliverable 22.1 'Persistent Identifiers Interoperability Framework' http://www.alliancepermanentaccess.org/index.php/about-aparsen/aparsen-deliverables/

[8] Research Data Alliance Recommendations in the field of Persistent Identifiers for digital objects https://rd-alliance.org/group/data-fabric-ig/wiki/persistent-identifier-bundle.html

[9] Persistent Identifier decision tree: http://www.ncdd.nl/en/pid

[10] Handle.Net https://www.handle.net

[11] The DOI® System http://www.doi.org

[12] ARK (Archival Resource Key) Identifiers https://confluence.ucop.edu/display/Curation/ARK

[13] Using National Bibliography Numbers as Uniform Resource Names, RFC 3188, J. Hakala https://tools.ietf.org/html/rfc3188

[14] Uniform Resource Names, https://datatracker.ietf.org/wg/urnbis/charter/

[15] Persistent URLs http://www.purlz.org and https://purl.oclc.org/docs/index.html

[16] Implementing Persistent Identifiers: Overview of concepts, guidelines and recommendations, Hans-Werner Hilse and Jochen Kothe, 2006 http://nbn-resolving.de/urn:nbn:de:gbv:7-isbn-90-6984-508-3-8

[17] Persistent identifiers in the Digital Preservation Handbook, Digital Preservation Coalition, http://handbook.dpconline.org/technical-solutions-and-tools/persistent-identifiers

[18] Persistent Identifiers (PIDs): recommendations for institutions, ATHENA WP3 Working Group, April 2011 http://www.athenaeurope.org/getFile.php?id=779

[19] Introduction to persistent identifiers, The THOR Project Knowledge Hub https://project-thor.readme.io/docs/introduction-to-persistent-identifiers

Workshop on Relational Database Preservation Standards and Tools

Luis Faria
KEEP SOLUTIONS
Rua Rosalvo de Almeida 5
4710 Braga, Portugal
lfaria@keep.pt

Marcel Büchler
Swiss Federal Archives
Archivstrasse 24
3003 Bern, Switzerland
marcel.buechler@bar.admin.ch

Kuldar Aas
National Archives of Estonia
J. Liivi 4
50409 Tartu, Estonia
kuldar.aas@ra.ee

ABSTRACT

This 3-hour workshop focuses on presenting the current state-of-the-art relational database preservation standards and tools used by major national archives and other institutions.

It presents **SIARD 2**, a new preservation format for relational databases. It also presents the current tools for harvesting information from live database management systems[1] into SIARD format and back, namely **SIARD Suite**[2] and the **Database Preservation Toolkit**[3]. Furthermore, two tools to access and view the information preserved in SIARD-files are presented: the **E-ARK database viewer**[4] and **SIARDexcerpt**[5].

The workshop includes live demonstration of the tools and prompts the participants to use them on their own laptops using the demonstration databases provided.

This workshop closely relates to a tutorial on relational database preservation guidelines and use cases, that focuses on the operational concerns of database preservation and relevant real-world use cases.

Keywords

Preservation; Archive; Relational Database; Ingest; Access; SIARD; Tools; E-ARK

1. INTRODUCTION

Databases are widely used to store and manage information. That is why archives have to ask themselves how they should preserve this type of information so that the databases will still be understandable in 20 or 50 years time. Furthermore, the database content needs to be stored independent of its specific Database Management System (DBMS), because a DBMS usually uses proprietary formats whose specifications are not freely accessible. In 2007 the Swiss Federal Archives (SFA) developed and standardised such an open format named SIARD (Software Independent Archiving of Relational Databases). Since 2008 SIARD has been actively used in the Swiss Federal Administration. Subsequently, the format has spread around the world, and is now used in many archives. In 2013 the SFA and the Swiss Coordination Agency for the Preservation of Electronic Files (KOST-CECO) specified the SIARD format as an eCH standard (Swiss E-Government Standards). working together with the E-ARK project (European Archival Records and Knowledge Preservation) the version 2.0 of the SIARD format was developed in 2015. To convert databases into SIARD files there are two existing tools: SIARD Suite (developed by the SFA) and the db-preservation-toolkit (developed by KEEP SOLUTIONS from Portugal).

2. OUTLINE

The workshop starts with a brief general introduction to the topic of database preservation, to familiarise the participants with the motivation and challenges of this topic.

In order to understand the possibilities and limitations of database preservation using the SIARD 2.0 some knowledge of the SIARD format is needed. Therefore we explain how the SIARD 2.0 format is based on four internationally recognised standards (XML, SQL:2008, UNICODE and ZIP64), and how SIARD files are structured.

Equipped with this theoretical background, workshop participants will be presented with some tools for database preservation. The Database Preservation Toolkit is open source software for conversion of live or backed-up databases into preservation formats such as SIARD. A demonstration database will be provided, so that the participants can experience themselves how to use the software. In order to make the preserved information accessible again, the E-ARK Database Viewer will be then demonstrated. The E-ARK Database Viewer allows the rapid ingest of SIARD files and provides a web application that allows on-line browsing and search of the database content. It also enables the printing and export of parts of the database into usable formats such as text and spreadsheet (e.g. Word and Excel).

After a break a second preservation tool is presented. SIARD Suite is a freeware reference implementation of the SIARD format developed by the SFA. Apart from database import/export, it also provides a basic viewer that can be used for user-friendly metadata enrichment of the SIARD files. Again, workshop participants are encouraged to try the software themselves on the demonstration database. Additionally, SIARDexcerpt, an open source viewer developed by the Swiss Coordination Agency for the Preservation of Electronic Files (KOST-CECO) is demonstrated. SIARDexcerpt allows users to search and select data directly from a standardised SIARD-file with a simple GUI and some configuration options. Neither a re-import into a database nor special know-how are necessary. The selected data can be represented in a human-readable form.

The workshop concludes with an open discussion and Q&A.

[1] e.g. Oracle, MySQL, Microsoft SQL Server

[2] https://www.bar.admin.ch/bar/en/home/archiving/tools/siard-suite.html

[3] http://www.database-preservation.com

[4] https://github.com/keeps/dbviewer

[5] https://github.com/Chlara/SIARDexcerpt

Table 1: Workshop timetable

Topic and duration	Presenter
Why preserve databases? Preservation strategies and problems (10 min)	Kuldar Aas
SIARD format 2.0 (20 min)	Marcel Büchler
Live demo and hands-on: Database Preservation Toolkit and E-ARK Database Viewer (60 min)	Luis Faria
Live demo and hands-on: Database Preservation Toolkit and E-ARK Database Viewer (60 min)	Luis Faria
Live demo and hands-on: SIARD Suite and SIARDexcerpt (60 min)	Marcel Büchler
Discussion and Q&A (30 min)	Luis Faria

See Table 1 for details on topics and timetable.

3. LEARNING OUTCOMES

At the end of the workshop the participants will have learned:

- The main problems and strategies to preserve relational databases
- Details on the SIARD relational database preservation format version 2.0
- How to use current state-of-the-art tools to preserve relational databases by harvesting them into SIARD format 2.0
- Some possibilities for how the preserved databases can be accessed, viewed and used

4. AUDIENCE AND REQUIREMENTS

This workshop targets digital preservation practitioners with some technical abilities and with interest in relational database preservation. Although not required, it is recommended to bring a laptop to test tools hands-on. The tools require Java version 7 or earlier to work. Access to demonstration databases will require wireless network access to be provided on-site.

5. PRESENTERS

Below are short biographies of the main presenters of this workshop:

Kuldar Aas

Deputy director of the Digital Archives of the National Archives of Estonia, working at the archives since 2002. Kuldar has participated in developing a set of national standards, requirements and guidelines in the areas of records management, metadata, transfer of electronic information to long-term archives, description and preservation of relational databases and Linked Open Data. He has also taken part in developing the Estonian proactive digital preservation and reuse environment at NAE. He is representing the national archives in Estonian semantic interoperability and Linked Data task forces.

Marcel Büchler

After finishing his studies in Computer Science, Marcel worked as a Database Engineer for a software development company. In 2015 he joined the Swiss Federal Archives where he is responsible for the SIARD format and the SIARD Suite.

Luis Faria

Innovation Director at KEEP SOLUTIONS, Portugal, and involved in European research projects focused on digital preservation, in particular SCAPE, 4C, E-ARK and VeraPDF (PREFORMA). Luis is taking his Ph.D. in digital preservation at the University of Minho. He qualified as a Systems and Informatics Engineer from the University of Minho. Luis was part of the original development team of RODA (Repository of Authentic Digital Objects) and has engaged in R&D tasks dedicated to systems design, platform development, format migration services and database preservation.

6. ACKNOWLEDGEMENTS

The SIARD format and SIARD Suite have been developed by the Swiss Federal Archives. SIARDexcerpt was developed by KOST-CECO, which is also involved in the eCH-standardisation of the SIARD format.

SIARD 2.0, the Database Preservation Toolkit and the E-ARK Database Viewer were partly developed under the auspices of E-ARK, an EC-funded pilot action project in the Competitiveness and Innovation Programme 2007-2013, Grant Agreement no. 620998 under the Policy Support Programme.

Understanding and Implementing PREMIS

Peter McKinney
National Library of New Zealand Te
Puna Mātauranga o Aotearoa
Cnr Molesworth & Aitken St
Wellington, New Zealand
+64 4 4623931
Peter.McKinney@dia.govt.nz

Eld Zierau
The Royal Library of Denmark
Søren Kierkegaards Plads 1
DK-1016 København K
+45 91324690
elzi@kb.dk

Evelyn McLellan
Artefactual Systems Inc.
Suite 201 – 301 Sixth Street
New Westminster, BC
Canada V3L 3A7
+1 604.527.2056
evelyn@artefactual.com

Angela Dappert
British Library
96 Euston Road
London
NW1 2DB
Angela.Dappert@gmail.com

ABSTRACT

This tutorial will provide participants with an introduction to the PREMIS Data Dictionary. It will give a basic overview of the standard and explore different models of implementation.

Keywords

Preservation strategies and workflows; Infrastructure, systems, and tools; Case studies, best practices and novel challenges; Training and education.

1. INTRODUCTION

The PREMIS Data Dictionary for Preservation Metadata is a specification that provides a key piece of infrastructure for digital preservation activities, playing a vital role in enabling the effective management, discovery, and re-usability of digital information. Preservation metadata provides provenance information, documents preservation activity, identifies technical features, and aids in verifying the authenticity of digital objects. PREMIS is a core set of metadata elements (called “semantic units”) recommended for use in all preservation repositories regardless of the type of materials archived, the type of institution, and the preservation strategies employed.

2. SUMMARY OF TUTORIAL

The PREMIS Data Dictionary was originally developed by the Preservation Metadata: Implementation Strategies (PREMIS) Working Group in 2005 and revised in 2008 and 2015. It is maintained by the PREMIS Editorial Committee and the PREMIS Maintenance Activity is managed by the Library of Congress.

We have seen a constant call for PREMIS to undertake tutorials such as this as more and more organisations come to grips with digital preservation. This tutorial provides an introduction to PREMIS and its data model and an examination of the semantic units in the Data Dictionary organized by the entities in the PREMIS data model, objects, events, agents and rights.

In addition it presents examples of PREMIS metadata and a discussion of implementation considerations, particularly using PREMIS in XML and with the Metadata Encoding and Transmission Standard (METS). It will include examples of implementation experiences through the institutional experience of the tutors.

The tutorial aims at developing and spreading awareness and knowledge about metadata to support the long term preservation of digital objects.

3. CONTENT OUTLINE

The draft outline for the tutorial is outlined below.

Introduction to PREMIS

- *Background (brief history and rationale of PREMIS)*
- *Benefits of implementing PREMIS*

PREMIS in detail

- *Core entities*
- *Simple examples to build familiarity*

Implementation

- *PREMIS in METS*
- *Case studies*
- *Support and the PREMIS community*
- *Conformance*

Next Steps

- *Round table discussion for institutional plans*

Wrap up

INTENDED AUDIENCE

The tutorial will benefit individuals and institutions interested in implementing PREMIS metadata for the long-term management and preservation of their digital information but who have limited experience in implementation. Potential audience includes cultural heritage operators, researchers and technology developers, professional educators, and others involved in management and preservation of digital resources.

4. EXPECTED LEARNING OUTCOMES

Participants will understand:

- What PREMIS is and why it exists;
- How PREMIS has changed across versions;
- The benefits of implementing PREMIS;
- The nature of the existing PREMIS community;
- The critical role PREMIS plays in the digital preservation community.

In addition, participants will get insight into:

- How PREMIS may be used in conjunction with METS;
- How different organisations implement PREMIS within their own repositories;
- The nature of conformance with PREMIS.

5. SHORT BIOGRAPHIES OF ORGANIZERS

Peter McKinney is the Policy Analyst for the Preservation, Research and Consultancy programme at the National Library of New Zealand Te Puna Mātauranga o Aotearoa. He currently serves as Chair of the PREMIS Editorial Committee.

Eld Zierau is member of the PREMIS Editorial Committee, since 2013. She is a digital preservation researcher and specialist, with a PhD from 2011 within digital preservation. Originally, she is a computer scientist, and has worked with almost all aspects of IT in private industries for 18 years, before starting in digital preservation in 2007. She has been working with many aspects of digital preservation, and she is involved as an architect or a consultant on major initiatives such a new digital repository including data modeling of metadata for preservation.

Evelyn McLellan graduated from the Master of Archival Studies program at the University of British Columbia, Canada, in 1997. She worked as an archivist and records manager for several organizations prior to joining Artefactual Systems in 2008. Evelyn started at Artefactual as the first ICA-AtoM Community Manager, then became the lead analyst for Archivematica, an open-source digital preservation system. In September 2013 she took on the role of President when Artefactual founder Peter Van Garderen stepped aside to work full-time on archives systems research. Evelyn has a long-standing interest in digital preservation and open technologies for archives and libraries. She has served as a co-investigator on the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) Project and as Adjunct Professor at the University of British Columbia's School of Library, Archival and Information Studies. She is currently a member of the PREMIS (Preservation Metadata Implementation Strategies) Editorial Committee.

Dr Angela Dappert is the Project Manager for the EU-cofunded THOR project (project-thor.eu) on linking researchers, data and publications through persistent identifiers. She has widely researched and published on digital repositories and preservation. She has consulted for archives and libraries on digital life cycle management and policies, led and conducted research in the EU-co-funded Planets, Scape, TIMBUS, and E-ARK projects, and applied digital preservation practice at the British Library through work on digital repository implementation, digital metadata standards, digital asset registration, digital asset ingest, preservation risk assessment, planning and characterization, and data carrier stabilization. She has applied her work towards preservation of research data and processes, software environments and eJournals, with an emphasis on interoperability and standardisation. Angela holds a Ph.D. in Digital Preservation, an M.Sc. in Medical Informatics and an M.Sc. in Computer Sciences. Angela serves on the PREMIS Editorial Committee and the Digital Preservation Programme Board of NRS.

Introduction to Fedora 4

David Wilcox
DuraSpace
9450 SW Gemini Drive #79059
Beaverton, OR 97008
1-607-216-4548
dwilcox@duraspace.org

Andrew Woods
DuraSpace
9450 SW Gemini Drive #79059
Beaverton, OR 97008
1-607-216-4548
awoods@duraspace.org

ABSTRACT

Fedora is a flexible, extensible, open source repository platform for managing, preserving, and providing access to digital content. Fedora is used in a wide variety of institutions including libraries, museums, archives, and government organizations. Fedora 4 introduces native linked data capabilities and a modular architecture based on well-documented APIs and ease of integration with existing applications. Both new and existing Fedora users will be interested in learning about and experiencing Fedora 4 features and functionality first-hand.

Attendees will be given pre-configured virtual machines that include Fedora 4 bundled with the Solr search application and a triplestore that they can install on their laptops and continue using after the workshop. These virtual machines will be used to participate in hands-on exercises that will give attendees a chance to experience Fedora 4 by following step-by-step instructions. Participants will learn how to create and manage content in Fedora 4 in accordance with linked data best practices. Finally, participants will learn how to search and run SPARQL queries against content in Fedora using the included Solr index and triplestore.

Keywords

Fedora, repository, linked data, open source.

1. OUTLINE

The tutorial will include three modules, each of which can be delivered in 1 hour.

1.1 Introduction and Feature Tour

This module will feature an introduction to Fedora generally, and Fedora 4 in particular, followed by an overview of the core and non-core Fedora 4 features. It will also include a primer on data modeling in Fedora 4, which will set the audience up for the next section.

1.2 Linked Data and LDP

The Fedora community is deeply invested in linked data best practices; this is exemplified by our alignment with the W3C Linked Data Platform recommendation in Fedora 4. This section will feature an introduction to linked data and LDP, with a particular focus on the way Fedora implements linked data. Attendees will have an opportunity to create and manage content according to linked data best practices using the Fedora 4 virtual machine.

1.3 Fedora 4 Integrations

Fedora 4 is fundamentally a middleware application – it is meant to be used in conjunction with other applications. This section will provide an overview of the most common integrations, such as Solr and triplestores. Attendees will learn how to use these tools to index and query content in Fedora.

2. DURATION

Half-day (3 hours)

3. AUDIENCE

This tutorial is intended to be an introduction to Fedora 4 - no prior experience with the platform is required. Repository managers and librarians will get the most out of this tutorial, though developers new to Fedora would likely also be interested.

4. OUTCOMES

Tutorial attendees will:

- Learn about the latest and greatest Fedora 4 features and functionality
- Discover new opportunities enabled by LDP and linked data
- Learn how to create and manage content in Fedora
- Understand how to index and query content in Fedora

5. PRESENTERS

David is the Product Manager for the Fedora project at DuraSpace. He sets the vision for Fedora and serves as strategic liaison to the steering committee, leadership group, members, service providers, and other stakeholders. David works together with the Fedora Technical Lead to oversee key project processes, and performs international outreach to institutions, government organizations, funding agencies, and others.

Andrew is a software engineer specializing in the coordination of open source, distributed development initiatives that focus on the preservation and access of digital cultural heritage. He has over a decade of experience advising, managing, and implementing projects across government and academics sectors. For the last six years, he has worked as a member of the DuraSpace team providing software development and community coordination of the DuraCloud and Fedora applications. Prior to joining the not-for-profit organization, DuraSpace, he worked as a software contractor on a number of Federal projects.

How can the UNESCO PERSIST Programme bring about Digital Sustainability through Legacy Software Services?

*Janet Delve, *David Anderson, †C. A. L. Lee, ‡Natasa Milic-Frayling

*University of Brighton
CRD, Grand Parade
Brighton, BN2 0JY, UK
{J.Delve, C.D.P.Anderson} @Brighton.ac.uk

†University of North Carolina
216 Lenoir Drive
Chapel Hill
North Carolina, USA
callee@ils.unc.edu

‡University of Nottingham
Jubilee Campus, Wollaton Road
Nottingham, NG8 1BB, UK
Natasa.Milic-Frayling@nottingham.ac.uk

ABSTRACT

This workshop will address the topic of sustained access to digital content by providing a legal framework and a technical platform for hosting and distributing functional legacy software. Both aspects represent key areas of the UNESCO PERSIST[1] [12] Programme that will focus on the preservation of the digital heritage under the UNESCO Memory of the World Programme.

The objective of the workshop is to engage Digital Preservation Practitioners, Memory Organizations, the ICT industry and policy makers in the discussion of (1) use cases for the international platforms of legacy software services, e.g., applications to preserving increasingly complex digital objects, (2) engagement models among the stakeholders that would lead to revenue streams and economically sustainable services, and (3) legal frameworks that ensure flexible use of legacy software in the far future, e.g., policies to guide life-cycle management of critical software and 'fair use of software' beyond its market lifespan.

The workshop fits most naturally into the innovative practices in digital preservation strand due to its pragmatic approach, but will also cover some research into digital preservation given the novel nature of the topic.

KEYWORDS

Digital Sustainability; Legacy Software Services (LSS); Software Preservation; Hardware museums; technical environment metadata; file formats; technical registry; emulation; virtualization; computing platform; virtual machine.

1. UNESCO "PERSIST" INITIATIVES

It is well recognized that digital information is difficult to preserve both in the short and the long term. Most storage media are short lived. Floppy disks and CD-ROMs were widely used for archiving digital content but are now outdated. Without a concerted effort to move the content to new storage media, content becomes inaccessible. However, even that measure is not sufficient. We may not be able to use the stored content because the software required to interpret the digital encoding cannot be run in the contemporary computing ecosystem. In other words, the software became obsolete and unusable and that, in turn, makes the legacy content inaccessible.

Software obsolescence is a side effect of the ongoing innovation in ICT industry. All software becomes outdated as vendors respond to market needs adding new features to existing products, or replacing them altogether. With diminished demand for previous versions of products, it becomes economically unfeasible to maintain them. Yet, without continuous updates the software becomes unusable and that, in turn, makes the content inaccessible in its original form. This is a key issue for the memory institutions who need to preserve digital content far longer than the life-span of software products and software vendors.

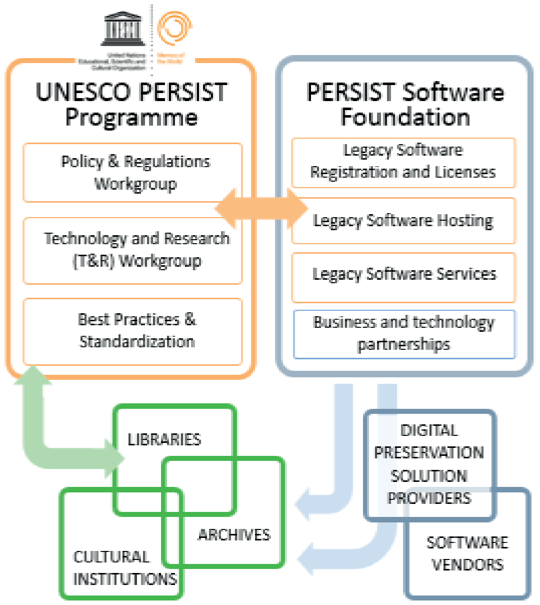

UNESCO PERSIST Programme
Policy & Regulations Workgroup
Technology and Research (T&R) Workgroup
Best Practices & Standardization
PERSIST Software Foundation
Legacy Software Registration and Licenses
Legacy Software Hosting
Legacy Software Services
Business and technology partnerships
LIBRARIES
ARCHIVES
CULTURAL INSTITUTIONS
DIGITAL PRESERVATION SOLUTION PROVIDERS
SOFTWARE VENDORS

Figure 1. Roles of the UNESCO PERSIST Programme and the PERSIST Software Foundation in securing sustainability of digital computation and digital content

One of the key mandates of the UNESCO is the preservation of world's heritage and unrestricted access to documentation to all. As cultural heritage is increasingly digital, UNESCO has recognized the global need to addressing the problem of digital obsolescence.

In December of 2013, UNESCO convened an international meeting in The Hague, the Netherlands, involving representatives from the ICT industry, governments and heritage institution to discuss opportunities to join forces and address the issues. In an open discussion of the business principles that guide the technology innovation, including input from Microsoft and International Chamber of Commerce, it has become clear that there are three essential aspects: legal, economic and technological issues. All three need to be addressed in parallel to enable long term access to digital legacy. Such a task requires thought leadership and coordination. For that reason, UNESCO established the UNESCO PERSIST Project comprising three taskforce focused on Technology, Content selection, and Policy.

In March 2016, UNESCO adopted PERSIST as its internal programme, as part of the well-established Memory of the World Programme, and structured it to include three Workgroups: Policy, Technology & Research (T&R), and Best Practices (see Figure 1).

1.1 Towards Sustainable Computation

Key to providing long term access to digital computation is reducing or eliminating the cost of software maintenance while retaining functionality. Virtualization/emulation technologies can shift the cost of software maintenance to the maintenance of virtual machines which can host a range of applications. Thus cost is spread across many applications and, potentially, many users.

Emulation is now a viable preservation strategy, supported by resources developed within publically funded research initiatives:

- Keeping Emulation Environments Portable (KEEP) project [10] produced an Emulation Framework [8], the Trustworthy Online Technical Environment Metadata (TOTEM) registry [11], and a media transfer knowledge base [9].
- bwFLA provided the DPC award-winning 'Emulation as a Service' [4, building on the TOTEM data model. This brings emulation within the technical grasp of memory institutions (legal issues are still to be addressed [1]).
- Digital Preservation Technical Registry (DPTR), underpinned by the TOTEM data model, is an ongoing effort carried out by the UoB (previously UoP), NLNZ, NLA, NASLA and NARA.
- The BitCurator set of tools [3] use digital forensic techniques to tackle media transfer and complex digital preservation tasks.
- The Preservation of Complex Objects Symposia (POCOS) project [7], concerned with preserving video games, digital art and archaeological 3D images, uses emulation as a key preservation approach, in addition to virtualization, software preservation [6, 5] and retaining hardware in computing museums [2].

1.2 Towards Policy and Legal Framework

In order to ensure a pervasive use of legacy software, it is key to establish a legal framework to manage the transition from commercial software licenses, designed to support vendors' business models, to the licenses appropriate for long term use of legacy software. Particularly important is the use of 'orphan' applications without legal guardians after the vendors stop trading.

1.3 Towards Economic Sustainability

Services that host and provide long term access to software need to generate sufficient value to users in order to generate revenue and sustain their operations. Thus, it is key to understand use scenarios that the services should support. The legacy software use is expected to be rare but important, particularly for computationally intense and interactive applications.

In order to provide effective services, UNESCO PERSIST Programme intends to incorporate a PERSIST Software Foundation, a non-for-profit legal entity with a formal partnership with the UNESCO but otherwise esponsible for its economic sustainability. Figure 1 outlines services that could form the core of the PERSIST Software platform and generate revenue streams.

1.4 Partnerships

Critical to the success of the UNESCO PERSIST Programme is the cooperation with the Digital Preservation community, ICT industry, solutions providers and the professional organizations that are already committed to the innovation and best practices, including ICA, IFLA, LIBER, etc.

2. WORKSHOP DISCUSSION POINTS

We will invite Workshop participants to engage in in-depth discussions of (1) use cases that could use Legacy Software Services (LSS), (2) engagement models with the platform that would lead to revenue streams and economically sustainable services, and (3) issues that the legal frameworks should cover in order to support the work of memory institutions. Among the discussion points it would be important to include:

- How could LSS complement current DP services?
- What specific services should LSS offer to the memory institutions and in what form (remote access, on premise, etc.)
- How will Open Source community collaborate with LSS?
- How will software vendors interact with the LSS?
- What would be role of Cloud infrastructure providers?

3. REFERENCES

[1] Anderson D. "The impact of European copyright legislation on digital preservation activity: lessons learned from legal studies commissioned by the KEEP project" in Preserving Complex Dig. Objects (Ed. Delve, J. & Anderson D.), 2014

[2] Bergmeyer W., Delve J., & Pinchbeck D. "Preserving games environments via TOTEM, KEEP and Bletchley Park" in Delve, J. & Anderson D. op.cit.

[3] BitCurator (http://www.bitcurator.net/)

[4] bwFLA (http://bw-fla.uni-freiburg.de/)

[5] Chue Hong N. "Digital preservation and curation: the danger of overlooking software." in Delve, J. & Anderson D. op.cit.

[6] Matthews B., Shaon A., & Conway E. "How do I know that I have preserved software?" in in Delve, J. & Anderson D.

[7] POCOS (http://www.cdpa.co.uk/POCOS

[8] The KEEP Emulation Framework (http://emuframework.sourceforge.net/)

[9] The KEEP Mediabase (http://mediabase.keep-totem.co.uk/)

[10] The KEEP Project (http://www.keep-project.eu/)

[11] TOTEM (http://www.keep-totem.co.uk/)

[12] UNESCO Digital Roadmap kicks off under the name PERSIST https://www.unesco.nl/en/node/2665

[1] PERSIST stands for Platform to Enhance the Sustainability of the Information Society Transglobally.

Building The Business Case And Funding Advocacy For Data Management Services

Paul Stokes
Jisc
One Castlepark Tower Hill
Bristol BS2 0JA
+44 (0)7595 056606
paul.stokes@jisc.ac.uk

William Kilbride
Digital Preservation Coalition
11 University Gardens
Glasgow G12 8QQ
+44 (0)141 330 4522
william.kilbride@dpconline.org

Neil Beagrie
Charles Beagrie Ltd
2 Helena Terrace
Salisbury SP1 3AN
+44 (0)1722 338482
neil@beagrie.com

ABSTRACT

In this hands on workshop we will address the tools, models and process used when building the business case for data management services including those relating to, research data, preservation, curation, discovery and access. We examine and test the use of existing tools with real world institutional problems and potential future tools/services.

Keywords

Research data management; Sustainability; 4C project; CESSDA SaW; Business case; Cost-benefit advocacy.

1. INTRODUCTION

In the current economic climate it is widely acknowledged that a robust business case is a prerequisite for a sustainable data management service. It is also becoming clear—particularly in the case of research data and the associated funder mandates—that data management needs to be considered holistically; all aspects of the data life cycle from the active creation phase through to preservation/curation (and ultimately disposal) affect the costs and benefits and need to be accounted for.

A number of tools and modelling techniques have emerged in recent years that allow practitioners to estimate costs and benefits which in turn help in the formulation of a business case. However, it is not clear if these are fit for purpose in this holistic context. Past 4C[1] work has shed some light on cost and benefit estimation techniques that are particularly suited to the post data publication phase, but these haven't been applied to the whole research data management (RDM) lifecycle. We are by no means sure if they can be applied in this fashion. And if they can't be used, nor do we know why in particular they may not be appropriate.

The CESSDA SaW project[2] is funded by the Horizon 2020 programme. Its principal objective is to develop the maturity of every national data archive service in Europe in a coherent and deliberate way towards the vision of a comprehensive, distributed and integrated social science data research infrastructure, facilitating access to social science data resources for researchers regardless of the location of either researcher or data. The funding advocacy toolkit being developed as part of the project will draw on a range of projects and studies looking at benefits, costs, return on investment and advocacy including inter alia 4C, Keeping Research Data Safe (KRDS), and a range of economic impact studies.

2. SCOPE

In the first part of the workshop we wish to explore which tools and models—in particular the CESSDA SaW project funding and cost-benefit advocacy toolkit and those tools and methodologies researched over the duration of the 4C project—might be applicable/appropriate when it comes to formulating a business case. It will be conducted with input from those who have put together business cases for data management services, both complete services and particular component services. We will explore their experiences, how they articulated their business cases and see if the lessons learned could benefit others. We will also address the problems of what it is that's stopping people from producing business cases and local barriers to progress.

In the second, more practical part of the workshop, in small groups we will work through institutional problems provided by attendees. We will test some of the current tools, identify the practical steps needed when using them to produce business cases and highlight any problems that might be encountered when using them in particular contexts. More experienced practitioners will be on hand to support participants in using the tools.

3. INTENDED FORMAT

Half day workshop

Session 1

- Introduction to the workshop
- Examination of the current tools with provocations from current users regarding their fitness for purpose
- Examples of real world business cases and lessons learned

Session 2

- Group work—practical application of current tools and methods to particular problems brought in by participants

Session 3

- Summing up

3.1 Planned outputs

After attending participants should have:

- A clear comprehension of the current understanding/ thinking around business cases for data management services
- Practical strategies for producing business cases with current tools in real world situations
- Understand the purpose of CESSDA SaW and the toolkit
- An indication of areas for further investigation with a view to the future provision of modified tools and services to address the end to end data management lifecycle

3.2 Speakers

The leads for the workshop will be William Kilbride (DPC), Neil Beagrie (CESSDA SaW) and Paul Stokes (Jisc). Other Jisc / 4C / CESSDA SAW project partners will contribute and, if possible, affiliate stakeholder organisations will also present.

3.3 Intended Audience

Practitioners, Managers and Funders—this has applicability at all levels and should be of practical, tactical and strategic interest.

The workshop is free.

3.4 Programme Strand

Aspects of this workshop straddle both the research and practice categories of the conference. However, it is intended to be primarily a practical workshop addressing the application of currently available tools and models to real world costing and business case scenarios. The identification of gaps and areas for further development addresses the research category.

It impacts upon various iPRES themes:

- Preservation strategies and workflows: preservation planning, access provision, risk analysis;
- Digital preservation frameworks: Digital preservation requirements and implications for the system lifecycle, business models, sustainability and economic viability.
- Infrastructure, systems, and tools: preservation resources

4. ACKNOWLEDGMENTS

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

The CESSDA SaW project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 674939

[1] http://4cproject.eu—4C is a recently completed FP7 European project—a Collaboration to Clarify the Cost of Curation

[2] http://cessda.net/CESSDA-Services/Projects/CESSDA-SaW

Towards Smarter Persistent Identifiers

Jonathan Clark
International DOI Foundation
Rembrandtlaan 12
1231 AC Loosdrecht
The Netherlands
+31654733572
jonathanmtclark@gmail.com

ABSTRACT

This workshop is a follow up to the Tutorial on Persistent Identifiers. The goal of the workshop is to discuss in detail key current issues and recent developments in persistent identifiers that are interesting for the Digital Preservation community

This workshop deals directly with new initiatives and capabilities of persistent identifiers.

Keywords

Persistent Identifiers; Smarter Persistent Identifiers; Multiple Resolution; Handle System; DOI; ARK; URN; URN:NBN.

1. INTRODUCTION

It is obvious that we need identifiers in a digital world to unambiguously determine what a resource is and where it can be found. However, over time it has become widely acknowledged that there is often great value in creating identifiers that are also persistent. As a result many different systems have emerged to address this need, so that a link to a resource can be guaranteed to work into the future [1].

Moreover, the scope of persistent identifiers has grown way beyond documents to include datasets, software, audiovisual content, people and other entities. Furthermore, innovative services have been developed alongside the identifiers. For instance, there is a service that allows funding agencies to track the impact of its funding across the academic domain [2].

A characteristic of the world today is that (content) domains are no longer discrete and independent. Researchers for instance want to make their datasets available and to link them to their formal publications. Libraries and archives are increasingly wanting to link out to data stored elsewhere rather than ingest everything themselves. As the domain entwine there is a growing realisation that the different persistent identifier systems must agree a framework for interoperability between the systems.

The European Commission under the Horizon 2020 programme had funded the THOR Project [3] to establish seamless integration between articles, data, and researchers. The objectives are to establish interoperability and integrate services across the partners. The goal is to make things easier and more open for researchers.

Most Persistent Identifiers resolve to a single resource but increasingly there is a need for multiple resolution where one entity to be resolved to multiple other pieces of data or entities [4].

We need smarter persistent identifiers. Identifiers that are machine readable and that do so much more than map to a single URL. We need identifiers that can work together to make things easier for everyone.

Smarter persistent identifiers have been discussed a lot in the IETF URNBIS working group [5]. One aspect of this is syntax: how should we attach resolution related requests to the identifier itself? Currently every system has a different solution to this problem, which is not an optimal situation.

The other issue is semantics: what kind of resolution services are there? Requesting provenance metadata is just one of the many, many services that potentially could be developed. Rights metadata has been discussed a lot (who has the copyright; is there a license; who is entitled to access the document; what can be done with it once it has been downloaded), and preservation metadata is another hot topic (what has happened to the document during archiving). What is clear is that HTTP is not sufficient; we need intelligent resolvers, which can pass requests to the correct servers in the right form. And in order to achieve this, our persistent identifiers must get a lot smarter than they currently are.

2. SCOPE

This workshop will present the latest developments in the Persistent Identifier world that are relevant for the iPRES community. This includes interoperability and (metadata) services such as multiple resolution.

3. TOPICS

The topics will depend to a great extent on the participants and their experiences and case studies. However, it is anticipated that at least the following topics will be covered in the workshop.

- The THOR project - establishing seamless integration between articles, data, and researchers
- Interoperability frameworks
- Multiple resolution
- Contextual resolution

4. INTENDED AUDIENCE

This workshop has a strong practitioner focus although it will cover only the most recent developments. It will be especially interesting for those working with Digital Archives and Digital Collections and all those who understand how important smarter persistent identifiers are. It will be an opportunity to share experiences and to discuss use cases in a highly interactive way.

5. EXPECTED LEARNING OUTCOMES

The aim of this workshop is to stimulate discussion and the exchange of ideas between the iPRES community and those involved in building smarter persistent identifiers. There is a need to connect the dots not only between the different systems but also between the different people involved.

6. WORKSHOP DESIGN

The audience is expected to contribute their experiences and case studies. In any event, there will be contributions from ORCID/DataCite (THOR), IDF, The Office of Publications of the European Union and RDA.

The workshop leader will be Jonathan Clark. Alongside his part-time work as Managing Agent for the DOI Foundation, Jonathan also designs, produces and runs innovative workshops for clients all over the world.

7. REFERENCES

[1] Persistent Identifiers: Considering the Options, Emma Tonkin, Ariadne Issue 56, 30th July 2008 http://www.ariadne.ac.uk/issue56/tonkin/

[2] Funding Data, a Crossref Service http://www.crossref.org/fundingdata/

[3] THOR: Technical and Human Infrastructure for Open Research https://project-thor.eu

[4] The DOI® Handbook, Chapter 3 ‘Resolution’ 2016 http://www.doi.org/doi_handbook/3_Resolution.html#3.8.4.3

[5] Uniform Resource Names, Revised (urnbis) https://datatracker.ietf.org/wg/urnbis/documents/

Acquiring and Providing Access to Born-Digital Materials with the BitCurator Environment and BitCurator Access Webtools

Christopher A. Lee
School of Information and Library Science
University of North Carolina
216 Lenoir Drive, CB #3360
1-(919)-966-3598
callee@ils.unc.edu

ABSTRACT

This tutorial will prepare participants to use the open-source BitCurator environment and BitCurator Access Webtools to acquire, process and provide access to born-digital materials. There will be a brief lecture and discussion that focuses on the motivation for using the tools and several foundational technical concepts. The remainder of the tutorial will be devoted to demonstration and hands-on exercises that feature specific tools and methods. Participants will learn how to mount media as read-only, create disk images, mount forensically packaged disk images, export individual files or entire directories from disk images, use specialized scripts to perform batch activities, generate and interpret Digital Forensics XML (DFXML), generate a variety of standard and customized reports (including PREMIS records), identify various forms of sensitive data within collections, and provide browser-based search and navigation of files and folders.

Categories and Subject Descriptors

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

General Terms

Provenance; Data Triage; Digital Forensics.

Keywords

Forensics; preservation; DFXML; metadata; privacy; collections; acquisition; web access

1. BITCURATOR PROJECT

The BitCurator Project, a collaborative effort led by the School of Information and Library Science at the University of North Carolina at Chapel Hill and Maryland Institute for Technology in the Humanities at the University of Maryland, is addressing two fundamental needs and opportunities for collecting institutions: (1) integrating digital forensics tools and methods into the workflows and collection management environments of libraries, archives and museums and (2) supporting properly mediated public access to forensically acquired data [4].

2. BITCURATOR ENVIRONMENT

We are developing and disseminating a suite of open source tools. These tools are being developed and tested in a Linux environment; the software on which they depend can readily be compiled for Windows environments (and in most cases are currently distributed as both source code and Windows binaries). We intend the majority of the development for BitCurator to support cross-platform use of the software. We are freely disseminating the software under an open source (GPL, Version 3) license. BitCurator provides users with two primary paths to integrate digital forensics tools and techniques into archival and library workflows.

First, the BitCurator software can be run as a ready-to-run Linux environment that can be used either as a virtual machine (VM) or installed as a host operating system. This environment is customized to provide users with graphic user interface (GUI)-based scripts that provide simplified access to common functions associated with handling media, including facilities to prevent inadvertent write-enabled mounting (software write-blocking).

Second, the BitCurator software can be run as a set of individual software tools, packages, support scripts, and documentation to reproduce full or partial functionality of the ready-to-run BitCurator environment. These include a software metapackage (.deb) file that replicates the software dependency tree on which software sources built for BitCurator rely; a set of software sources and supporting environmental scripts developed by the BitCurator team and made publicly available at via our GitHub repository (links at http://wiki.bitcurator.net); and all other third-party open source digital forensics software included in the BitCurator environment.

3. BITCURATOR ACCESS WEBTOOLS

The BitCurator Access project has developed BCA Webtools, which is a suite of software that allows users to browse a wide range of file systems contained within disk images using a web browser. It is intended to support access requirements in libraries, archives, and museums preserving born-digital materials extracted from source media as raw or forensically-packaged disk images.

BCA Webtools uses open source libraries and toolkits including The Sleuth Kit, PyTSK, and the Flask web microservices framework. It uses PyLucene along with format-specific text-extraction tools to index the contents of files contained in disk images, allowing users to search for relevant content without individually inspecting files. BCA Webtools is distributed with a simple build script that deploys it as a Vagrant virtual machine running the web service.

4. TUTORIAL FORMAT

There will be a brief lecture and discussion that focuses on the motivation for using the tools and several foundational technical concepts. The remainder of the tutorial will be devoted to demonstration and hands-on exercises that demonstrate specific tools and methods.

5. INTENDED AUDIENCE

This tutorial should be of interest to information professionals who are responsible for acquiring or transferring collections of digital materials, particularly those that are received on removable media. Another intended audience is individuals involved in digital preservation research, development and IT management, who will learn how data generated within the BitCurator environment can complement and potentially be integrated with data generated by other tools and systems.

6. EXPECTED LEARNING OUTCOMES

This tutorial will prepare participants to use the open-source BitCurator environment to acquire and process born-digital data. Tools that BitCurator is incorporating include Guymager, a program for capturing disk images; bulk extractor, for extracting features of interest from disk images (including private and individually identifying information); fiwalk, for generating Digital Forensics XML (DFXML) output describing filesystem hierarchies contained on disk images; The Sleuth Kit (TSK), for viewing, identifying and extraction information from disk images; Nautilus scripts to automate the actions of command-line forensics utilities through the Ubuntu desktop browser; and sdhash, a fuzzing hashing application that can find partial matches between similar files. For further information about several of these tools, see [1,2,3,5].

Upon completion of this tutorial, participants should understand several of the major motivations and uses cases for applying the BitCurator environment. They will also know how to perform the following tasks:

- mount media as read-only
- create disk images, mount forensically packaged disk images
- export individual files or entire directories from disk images
- use specialized scripts to perform batch activities
- generate and interpret Digital Forensics XML (DFXML) generate a variety of standard and customized reports (including PREMIS records)
- identify various forms of sensitive data within collections
- provide browser-based search and navigation of files and folders.

Participants will also become aware of the resources that are available for learning more about the software and engage with other users after completion of the tutorial.

7. INSTRUCTOR BIOGRAPHY

Christopher (Cal) Lee is Associate Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He teaches graduate and continuing education courses in archival administration, records management, digital curation, and information technology for managing digital collections. His research focuses on curation of digital collections and stewardship of personal digital archives. Cal is PI for the BitCurator project and editor of *I, Digital: Personal Collections in the Digital Era*.

8. ACKNOWLEDGMENTS

Development of the BitCurator environment and BitCurator Access Webtools have been supported by the Andrew W. Mellon Foundation.

9. REFERENCES

[1] Cohen, M., Garfinkel, S., and Schatz, B. 2009. Extending the Advanced Forensic Format to Accommodate Multiple Data Sources, Logical Evidence, Arbitrary Information and Forensic Workflow. *Digital Investigation* 6 (2009), S57-S68.

[2] Garfinkel, S. Digital Forensics XML and the DFXML Toolset. *Digital Investigation* 8 (2012), 161-174.

[3] Garfinkel, S.L. Providing Cryptographic Security and Evidentiary Chain-of-Custody with the Advanced Forensic Format, Library, and Tools. *International Journal of Digital Crime and Forensics* 1, 1 (2009), 1-28;

[4] Lee, C.A., Kirschenbaum, M.G., Chassanoff, A., Olsen, P., and Woods, K. BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions. *D-Lib Magazine* 18, 5/6 (May/June 2012).

[5] Roussev, V. An Evaluation of Forensic Similarity Hashes. *Digital Investigation* 8 (2011), S34-S41.

PREMIS Implementation Fair

Peter McKinney
National Library of New Zealand Te Puna Mātauranga o Aotearoa
Cnr Molesworth & Aitken St
Wellington, New Zealand
+64 4 4623931
Peter.McKinney@dia.govt.nz

Eld Zierau
The Royal Library of Denmark
Søren Kierkegaards Plads 1
DK-1016 København K
+45 91324690
elzi@kb.dk

Evelyn McLellan
Artefactual Systems Inc.
Suite 201 – 301 Sixth Street
New Westminster, BC
Canada V3L 3A7
+1 604.527.2056
evelyn@artefactual.com

Angela Dappert
British Library
96 Euston Road
London
NW1 2DB
Angela.Dappert@gmail.com

ABSTRACT

This workshop provides an overview of the PREMIS Data Dictionary for Preservation Metadata, a standard addressing the information you need to know to preserve digital content in a repository. It includes an update on current PREMIS initiatives and reports from the preservation community on implementation of the standard in various systems or contexts.

Keywords

Preservation metadata, Preservation repository implementation, Data dictionary

1. INTRODUCTION

The PREMIS Implementation Fair Workshop is one of a series of events organized by the PREMIS Editorial Committee and that has been held in conjunction with previous iPRES conferences.

At iPRES 2016, the workshop will give the audience a chance to understand updates in the PREMIS data dictionary and give implementers, and potential implementers, of the *PREMIS Data Dictionary for Preservation Metadata* an opportunity to discuss topics of common interest and find out about latest developments.

2. OUTLINE OF WORKSHOP CONTENT

2.1 Overview of Updates to the PREMIS Data Dictionary

The *PREMIS Data Dictionary for Preservation Metadata* [1] is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. Developed by an international team of experts, PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into a number of commercial and open-source digital preservation tools and systems. This session provides an overview of the PREMIS Data Model (which was recently revised) and of the types of information specified to support the digital preservation process. Included will be a summary of the changes in version 3.0, which includes enhanced ability to describe intellectual objects and technical environments within the PREMIS context.

2.2 Implementation Reports

Implementation reports will be solicited from the PREMIS Implementers community. The workshop is one of the only times of the year that implementers can come together to show and discuss their implementations. They are crucial for not only fostering a sense of community, but also for institutions to get direct feedback on critical questions and challenges in their digital preservation programmes.

3. WORKSHOP SERIES

If accepted, the PREMIS Implementation Fair at iPRES 2016 would be the seventh in a series that has been held in conjunction with iPRES since 2009. These events are intended to highlight PREMIS activities, discuss issues concerning implementation, and provide a forum for implementers to discuss their activities, issues and solutions. Because this is a rapidly changing area, it is important to provide continuous updates. iPRES is the primary forum for this conversation between PREMIS and the user community.

4. INTENDED AUDIENCE

The workshop is designed for those involved in selecting, designing, or planning a preservation project or repository using preservation metadata. This includes digital preservation practitioners (digital librarians and archivists, digital curators, repository managers and those with a responsibility for or an interest in preservation workflows and systems) and experts of digital preservation metadata and preservation risk assessment.

5. PROCESS FOR SOLICITING CONTRIBUTIONS

Contributions will be solicited from the PREMIS Implementers' Group via its discussion list (pig@loc.gov). The PREMIS Editorial Committee will review all requests. If the workshop proposal is approved, a call will be sent for contributions to the implementation portion and the deadline will be within a month.

6. SHORT BIOGRAPHIES OF ORGANIZERS

Peter McKinney is the Policy Analyst for the Preservation, Research and Consultancy programme at the National Library of New Zealand Te Puna Mātauranga o Aotearoa. He currently serves as Chair of the PREMIS Editorial Committee.

Eld Zierau is member of the PREMIS Editorial Committee, since 2013. She is a digital preservation researcher and specialist, with a PhD from 2011 within digital preservation. Originally, she is a computer scientist, and has worked with almost all aspects of IT in private industries for 18 years, before starting in digital preservation in 2007. She has been working with many aspects of digital preservation, and she is involved as an architect or a consultant on major initiatives such a new digital repository including data modeling of metadata for preservation.

Evelyn McLellan graduated from the Master of Archival Studies program at the University of British Columbia, Canada, in 1997. She worked as an archivist and records manager for several organizations prior to joining Artefactual Systems in 2008. Evelyn started at Artefactual as the first ICA-AtoM Community Manager, then became the lead analyst for Archivematica, an open-source digital preservation system. In September 2013 she took on the role of President when Artefactual founder Peter Van Garderen stepped aside to work full-time on archives systems research. Evelyn has a long-standing interest in digital preservation and open technologies for archives and libraries. She has served as a co-investigator on the International Research on Permanent Authentic Records in Electronic Systems (InterPARES) Project and as Adjunct Professor at the University of British Columbia's School of Library, Archival and Information Studies. She is currently a member of the PREMIS (Preservation Metadata Implementation Strategies) Editorial Committee.

Dr Angela Dappert is the Project Manager for the EU-cofunded THOR project (project-thor.eu) on linking researchers, data and publications through persistent identifiers. She has widely researched and published on digital repositories and preservation. She has consulted for archives and libraries on digital life cycle management and policies, led and conducted research in the EU-co-funded Planets, Scape, TIMBUS, and E-ARK projects, and applied digital preservation practice at the British Library through work on digital repository implementation, digital metadata standards, digital asset registration, digital asset ingest, preservation risk assessment, planning and characterization, and data carrier stabilization. She has applied her work towards preservation of research data and processes, software environments and eJournals, with an emphasis on interoperability and standardisation. Angela holds a Ph.D. in Digital Preservation, an M.Sc. in Medical Informatics and an M.Sc. in Computer Sciences. Angela serves on the PREMIS Editorial Committee and the Digital Preservation Programme Board of NRS.

7. REFERENCES

[1] PREMIS Editorial Committee. 2015. *PREMIS Data Dictionary for Preservation Metadata.*

Tutorial on Relational Database Preservation: Operational Issues and Use Cases

Krystyna W. Ohnesorge
Swiss Federal Archives
Archivstrasse 24
3003 Bern, Switzerland
Tel. +41 58 464 58 27
krystyna.ohnesorge@bar.admin.ch

Marcel Büchler
Swiss Federal Archives
Archivstrasse 24
3003 Bern, Switzerland
Tel. +41 58 469 77 32
marcel.buechler@bar.admin.ch

Anders Bo Nielsen
Danish National Archives
Rigsdagsgaarden 9
DK-1218 Copenhagen K
+45 41 71 72 35
abn@sa.dk

Phillip Mike Tømmerholt
Danish National Archives
Rigsdagsgaarden 9
DK-1218 Copenhagen K
+45 41 71 72 20
pmt@sa.dk

Zoltan Lux
National Archives of Hungary
Bécsi kapu tér 2-4
1250 Budapest, Hungary
+36 1 437 0660
lux.zoltan@mnl.gov.hu

Janet Delve
University of Brighton
CRD, Grand Parade
Brighton, BN2 0JY, UK
+44 1273 641820
J.Delve@Brighton.ac.uk

Kuldar Aas
National Archives of Estonia
J. Liivi 4
50409 Tartu, Estonia
+372 7387 543
kuldar.aas@ra.ee

ABSTRACT

This 3-hour tutorial focuses on the practical problems in ingesting, preserving and reusing content maintained in relational databases. The tutorial is based on practical experiences from European national archives and provides an outlook into the future of database preservation based on the work undertaken in collaboration by the EC-funded E-ARK project[1] and the Swiss Federal Archives[2].

This tutorial relates closely to the workshop: "Relational database preservation standards and tools" which provides hands-on experience on the SIARD database preservation format and appropriate software tools.

Keywords

Relational database, database preservation, SIARD2, E-ARK

1. INTRODUCTION

With the introduction of paperless offices, more information than ever is being created and managed in digital form, often within information systems which internally rely on database management platforms and store the information in a structured and relational form.

Preserving relational databases and providing long-term access to these is certainly a complex task. Database preservation methods cannot concentrate only on the preservation of the data itself, but must also address multiple administrative and technical issues to provide a complete solution, satisfying both data providers and users.

Technically the *de facto* standard for preserving relational databases is the open SIARD format. The format was developed in 2007 by the Swiss Federal Archives (SFA) and has since then been actively used in the Swiss Federal Administration and also internationally. However, the administrative and procedural regimes, as well as the practical implementation of the standard, can vary to a large degree due to local legislation, best-practices and end-user needs.

2. OUTLINE

The tutorial starts with an overview of the problems and challenges in preserving relational databases. Following this introduction we present three national use cases (Switzerland, Denmark and Hungary) which highlight specific national aspects and best-practices. Finally, the tutorial will present database de-normalization and data mining techniques which are being currently researched and applied within the E-ARK Project.

Throughout the tutorial participants will have the possibility to contribute to an open discussion on the issues and solutions in relational database preservation.

Table 1: Tutorial overview

Topic and duration	*Presenter*
Introduction to preserving relational databases (30 min)	Kuldar Aas
National Use Case: Switzerland (30 min)	Marcel Büchler, Krystyna Ohnesorge
National Use Case: Denmark (30 min)	Anders Bo Nielsen, Alex Thirifays
National Use Case: Hungary (30 min)	Zoltan Lux
Advanced Database Archiving Scenarios (30 min)	Janet Delve
Open discussion (30 min)	

[1] http://www.eark-project.eu

[2] http://www.bar.admin.ch

2.1 Introduction to Preserving Relational Databases

This presentation sets the scene for the rest of the tutorial by highlighting some of the main issues in database preservation:

- How to convince data holders about the importance of database preservation?
- Which preservation method to use (i.e. migration vs emulation)?
- How to minimize the amount of work needed to be undertaken by system developers, data owners and archives especially during pre-ingest and ingest?
- How to ensure that the archiving process does not hinder the original data owner in providing their own services?
- How to ensure that the archived database is appropriately accessible to the users?

2.2 National Use Cases

2.2.1 Swiss Federal Archives

The Swiss Federal Archives (SFA) have been actively dealing with preserving relational databases for more than a decade. Most notably, they took charge in the early 2000's to develop the original SIARD format and the accompanying SIARD Suite software tools.

In the Swiss Federal Administration, there is a relatively large amount of freedom and flexibility in the DB design, therefore a variety of DB-models is encountered. In this use case SFA will focus on the coordination with data owners and explain the steps which need to be made in applying the SIARD approach in agencies.

2.2.2 Danish National Archives

The Danish National Archives (DNA) already started archiving government databases in the 1970's and has by now archived more than 4,000 databases. Over the decades they have gathered widespread experience in both administrative and technical issues.

In this use case DNA will focus on the most common issues which they have encountered while implementing their database preservation regime across the Danish public sector, and will provide an overview of the most common issues encountered during the ingest processes.

2.2.3 Hungarian National Archives

The use case of the National Archives of Hungary (NAH) will concentrate on the access and use of preserved databases. The specific problem addressed is the access to databases which originally have a complex data structure and are therefore impossible to be reused without in-depth knowledge about data modelling and structures.

The use case will demonstrate how to simplify archival access by de-normalizing the original database during ingest, and reusing it with the help of the Oracle Business Intelligence (Oracle BI)[3] platform and APEX[4] application.

2.3 Advanced Database Archiving Scenarios

As the last presentation the tutorial features an overview of the work undertaken in the E-ARK project on advanced database archiving and reuse.

More explicitly the presentation will cover the solutions for creating de-normalized AIPs ready for dimensional analysis via Online Analytical Processing (OLAP) and other tools. Essentially this is adding value to the AIPs by making them more discovery-ready, allowing complex analysis to be carried out on data within an archive, or even across several archives from different countries.

The main aim of the presentation is to discuss whether such an approach is possible to be automated to an extent that it would be applicable in preservation institutions with limited technical knowledge about relevant tools and methods.

3. AUDIENCE AND OUTCOMES

The tutorial targets equally preservation managers and specialists who need to establish routines for ingesting, preserving and reusing relational databases.

Participants will gain a broad overview of the problems and solutions which have been set up across Europe.

4. ACKNOWLEDGEMENTS

Part of the tutorial is presenting on work which has been undertaken within E-ARK, an EC-funded pilot action project in the Competiveness and Innovation Programme 2007-2013, Grant Agreement no. 620998 under the Policy Support Programme.

[3] https://www.oracle.com/solutions/business-analytics/business-intelligence/index.html

[4] https://apex.oracle.com/en/

Sharing, Using and Re-using Format Assessments

Andrea Goethals
Harvard Library
90 Mt. Auburn St.
Cambridge, MA 01970 USA
+1-617-495-3724
andrea_goethals@harvard.edu

Kate Murray
Library of Congress
101 Independence Ave
Washington, DC 20540 USA
+1-202-707-4894
kmur@loc.gov

Michael Day
The British Library
96 Euston Road
London NW1 2DP
+44 330 333 1144 ext. 3364
Michael.Day@bl.uk

Kevin L. De Vorsey
National Archives and Records Administration
One Bowling Green, Room 450
New York, NY 10004 USA
+1-212-401-1631
Kevin.devorsey@nara.gov

Jay Gattuso
National Library of New Zealand Te Puna Mātauranga o Aotearoa
Cnr Molesworth & Aitken St
Wellington, New Zealand
+64 4 4743064
Jay.Gattuso@dia.govt.nz

Paul Wheatley
Digital Preservation Coalition
37 Tanner Row
York YO1 6WP
+44 0 1904 601871
paul@dpconline.org

ABSTRACT

Many cultural heritage institutions with digital collections have performed assessments of file formats to inform decision making for a wide range of activities. The development of digitization standards and transfer requirements for depositors, the selection of storage and access platforms, and preservation planning and the decisions to use emulation or migration as a preservation approach all benefit from the assessment of file formats for their appropriateness in these different contexts. This workshop will bring together several institutions who have created format assessments, together with institutions who already are or could potentially reuse this work to inform their own institutional policies and preservation planning. The workshop will start with short presentations to expose the assessment work that has been done, followed by a discussion of how they are being used, or could be used, and possibilities for more effectively sharing these resources across institutions for the benefit of the digital preservation community.

Keywords

File Format Assessments; Recommended File Formats; Preservation Formats

1. FORMAT ASSESSMENTS

Several cultural heritage institutions responsible for preserving digital content have generated what could be called format assessments. While the specific workflow, criteria and artifacts created have tended to be institution-specific, and have even varied over time within the same institutions; at a high level all of these format assessments could be defined as the detailed documentation of the properties of a format to gain insight into the format's suitability to fill a repository function, e.g. as an archival format, as a transfer format, as an access format. They are used to support decisions related to content that may come into the repository, or that is already under management in the repository.

The institutions writing format assessments create them for various reasons. Some create them to inform policy and related guidelines for their preservation repository, for example as the basis for guidelines for content creators, or to restrict the formats accepted into their repository. Other institutions create them to inform the broader digital preservation community, potentially as the basis for best or at least good practices. Still others create them to make decisions about formats to select for normalization or migration targets, or to identify formats that might be at risk of obsolescence within their repository. Because of the diversity of the reasons for format assessments, among institutions creating them the methods used and the results are divergent. A key difference is how a "format" is defined or scoped in the assessment. Some, like those written for the Library of Congress' Sustainability of Digital Formats Web site [1] are very granular - there are eight different variations of the JPEG2000 JP2 format described on the site. In contrast the assessments written by the British Library [2] are not as granular, e.g. there is a single assessment covering the JPEG2000 JP2 format. The assessments created by Harvard Library [3] are also less granular than the Library of Congress' but include associated assessments of metadata and tools specific to formats.

Despite the differences in how and why format assessments are created, institutions of all types within the digital preservation community could benefit from the broader sharing of these assessments. There is a great deal of time, effort and resources that go into preparing format assessments. Leveraging the research and findings already done by other institutions will allow institutions to focus efforts on work not already done. In addition, institutions that do not have the resources to do their own format assessments are still able to write needed preservation policies and conduct preservation planning by reusing the work that has already been done. Lastly, exposing this work to more eyes in the community should lead to more discussion and constructive feedback about formats and their suitability for preservation that can lead to establishing community-accepted best practices in this area.

2. WORKSHOP STRUCTURE

The authors of this proposal represent institutions who are actively engaged in format assessments and wish to make this work discoverable and useful as a resource to the digital preservation community. The workshop will begin with brief descriptions from each of the authors on their role related to format assessments. The Library of Congress (LC) will present a case history perspective of its detailed PDF [4] assessments on the Sustainability of Digital Formats website to demonstrate the granularity of LC's assessments but also real life application of their usefulness. Harvard Library (HL) will describe the format and related metadata and tool assessments it has been doing as preparation for supporting new formats in HL's preservation repository. The British Library will outline the evolution of its file format assessment activities in the context of the development of a preservation planning capacity based on a deeper understanding of the Library's collections and preservation priorities. The National Archives and Records Administration (NARA) will describe the assessment methodology it undertakes to determine file formats that are appropriate for use by Federal agencies transferring permanent electronic records. The National Library of New Zealand (NLNZ) will explore the relationship between format identification and format assessments. This will draw on a decade of experience and will discuss the boundaries between them, the institutional mandates that inform their use and value, and point to some directions on how this information is best linked and shared.

The presentations will be followed by an informal discussion to explore the following topics:

- Use of format assessments
  - Are they being used by institutions that did not create them, and for what purposes?
  - Is there any additional information or infrastructure needed to make them more discoverable, interpreted as intended by the creators, or more useful to other institutions?
- Central discovery platform for format assessments
  - Is it desirable for the community to have a central portal for format assessments, and if so, how would this be maintained?
  - What would be the challenges to implementing this and what are some ideas for resolving them?

The last part of the workshop will focus on identifying concrete next steps following the workshop, including but not limited to an update of this paper after the workshop.

3. REFERENCES

[1] Sustainability of Digital Formats: Planning for Library of Congress Collections http://www.digitalpreservation.gov/formats/

[2] File Formats Assessments – wiki.dpconline.org http://wiki.dpconline.org/index.php?title=File_Formats_Assessments

[3] Format Assessments – Harvard Library Digital Preservation – Harvard Wiki https://wiki.harvard.edu/confluence/display/digitalpreservation/Format+Assessments

[4] PDF (Portable Focument Format) Family http://www.digitalpreservation.gov/formats/fdd/fdd000030.shtml

Active Data Management Planning: chances and challenges

Felix Engel
FernUniversität in Hagen, Germany
Felix.Engel@FernUni-Hagen.de

Heike Görzig
FernUniversität in Hagen, Germany
Heike.Goerzig@FernUni-Hagen.de

Rob Baxter
EPCC, University of Edinburgh
r.baxter@epcc.ed.ac.uk

Helen Glaves
British Geological Survey, United Kingdom
hmg@bgs.ac.uk

Simon Waddington
King's College London
United Kingdom
simon.waddington@kcl.ac.uk

Matthias Hemmje
FernUniversität in Hagen, Germany
Matthias.Hemmje@fernUni-Hagen.de

ABSTRACT

There is an increasingly urgent need to ensure the fullest possible preservation of research findings, both for proper validation of the research in question and to enable its comprehensive reuse. Furthermore, research undertakings are expensive and the return on investment needs to be secured by research funding agencies and public funding bodies through proper management of the knowledge that is required for the effective long-term reuse of results. Awareness of these facts leads to increasingly stringent regulations from funding agencies, which seek to enforce compliance of research data with specific policies. Hence, funding agencies are beginning to make *Data Management Plans* (DMP) increasingly mandatory before they will fund a research undertaking. In general, a DMP is a full text document that elaborates how the research data is handled, both during and after the project lifetime. In fact, a DMP includes policies for data management on different levels, as e.g. required by a formal framework the research has to comply with, as well as managerial parameters or policies that directly address the data management system level. Nevertheless, besides the pure establishment of policies, funders and researchers have further requirements concerning active aspects of data management, as e.g.: the continuous adoption of DMPs and its execution throughout the research lifecycle or to preserve the knowledge created during research that is needed for comprehensive later research validation. Because of the complexity of these and further requirements, management support is under discussion in various research communities. Within the international Research Data Alliance, these aspects are discussed within the so called Active DMP Interest Group (ADMP IG, cf. [1]).

This workshop will consider the outcomes of the next RDA ADMP IG workshop (cf. [2]) to discuss additional ADMP related topics and will address further open research questions around ADMPs, with a special focus on continuous adoption of DMPs and automation support. Hence, the aim of this workshop is to identify on base of the submitted contributions and the conclusion of the discussion during the workshop the recent obstacles that prevent the realization of ADMPs and how those could be addressed. The outcome of this workshop will be the preparation of a roadmap towards a reference framework of ADMP management and automation.

KEYWORDS

Active Data Management Planning, Policies.

1. SUBJECT

"*Innovation has been placed at the heart of the EU's strategy to create growth and jobs*". As a consequence the European Commission (EC) has encouraged EU countries to invest 3% of their GDP in R&D (cf. [3]). Nevertheless, numerous recent investigations point out that proper validation of research is not a matter of course. The benefits of expensive research are uncertain, because proper reuse is not free of doubt.

To secure the return on investment research funding agencies and public funding bodies are beginning to make proper data management planning of funded research a mandatory part within a project proposal. In general, a DMP is a full text document that describes how the research data is handled both during and after the project. The creation and management of these documents is technically supported by means of services like *DMPOnline* that is offered through the Digital Curation Centre (cf. [4]) or the *DMPTool* that is offered through the California Digital Library (cf. [5]). These services provide funder-specific DMP templates, guiding a user through all the required DMP declarations. But besides the specification of a DMP as a document, further requirements exist from funder and data manager perspective to consider dynamic aspects of a DMP, as a "living document" that is updated continuously and supports automating the policy enforcement while the project progresses. In fact, these living documents would be greatly supported through the existence of tools that would help the researchers in providing the additional information required throughout the lifecycle of the data. This need arises especially, because at the proposal stage some elements of the metadata and other information relating to the datasets are limited, as e.g. data formats or likely data volumes. Additionally, important it is to enable as well the comprehensive validation of DMP compliance, the monitoring of managed research data or to capture further knowledge that comes into existence along the research undertaking, required for comprehensive later research reuse. Several of mentioned aspects are already under discussion within the RDA, as so called *Active DMPs* (ADMP, cf. [1]). But, while service offers exists to support the creation and management of DMPs, further research is necessary to support as well these active aspects of data management and execution. An initial investigation towards the automation of ADMPs has been investigated in the RDA Practical Policy WG (cf. [6]). This WG discusses so called *actionable rules* that are derived from exiting DMPs, formalized and enforced through the application of the *integrated Rule Oriented Data System* (iRODS, cf. [7]) rule language.

2. SCOPE

Various actors participate in the process of DMP specification and execution, influencing research data management throughout the whole data lifecycle in various dimensions (cf. [8]). The **Formal Dimension** of DMPs is covered by the funding agencies' Grant Agreements (GA), corresponding legal requirements. The GAs usually provide the contractual framework for DMP; it specifies what the DMP has to accomplish and to comply with. Corresponding laws and regulations therefore provide the legal, regulatory and consequent policy building framework. To comply with the requirements and challenges created by the analysis of this *formal layer*, an RDM work plan is developed in the **Managerial Dimension** of DMP. The RDM work plan describes the RDM scenario that has to be created in order to comply with the DMP requirements and challenges and their corresponding representation schema set up by the analysis of the formal dimension. This RDM work plan includes strategic and organizational aspects, concrete activities, and deliverables. In the RDM work plan sequences of activities and their dependencies are formulated; thus the implementation of the DMP is based on this RDM work plan. The data producers who are, e.g., software developers and researchers on the project, form the **Operative Dimension** of the DMP. Tasks and activities listed in the work plan are executed by them, thereby producing and using the data to be archived and preserved for effective later reuse. Effectively, these three-dimension details the various stakeholders and requirement that are part of the DMP creation and are involved in its development and execution through its lifecycle. But, while the initial creation of DMPs is already supported by means of existing software applications, to date, the consideration and realization of dynamic aspects, spanning all involved actors in data management planning and its execution are rarely addressed within recent discussion and research. Hence, to get a better understanding about the incidents that actually prevent the consideration of the dynamics nature of DMPs, this workshop aims to discuss on the one hand those aspects that actually prevent dynamics in the management and realization of DMPs and on the other hand how these aspects could be addressed through the provision of software applications that enable automation in data management planning and DMP realization process.

3. PROGRAMME STRAND AND CONTENT

The overall scope and goal of the workshop is to bring together academic and industrial researchers and practitioners to discuss a common roadmap to support the acceleration of ADMP-related research activities and to achieve a common understanding of the overall requirements. This roadmap can be used to inform, influence and disseminate ideas to funders, the wider research community, and the general public. Thus, this workshop will maximize the benefit of DMPs for a range of stakeholders. To address this objective, the workshop will bring up and discuss open research questions around active data management planning and automated execution, with a special focus on supporting the treatment of DMPs as living documents, updated and automated where possible throughout the single life cycle phases.

The workshop will start with the presentation of accepted contributions and will be followed by a discussion about identified open issues and preparation of a roadmap that will address the realization of a reference framework towards ADMP management and automation.

4. ORGANISATIONAL

Workshop chair: Matthias Hemmje

Co-chairs: Felix Engel, Heike Görzig, Simon Waddington, Helen Glaves.

5. REFERENCES

[1] RDA, ADMP IG. URL: https://rd-alliance.org/group/active-data-management-plans/case-statement/active-data-management-plans-ig.html [Accessed 27.042016]

[2] RDA Workshop on ADMP. URL: https://indico.cern.ch/event/520120/ [Accessed 27.042016]

[3] EUROPA - Topics of the European Union - Research & innovation. [Online]. Available: http://europa.eu/pol/rd/. [Accessed: 17-Mar-2016].

[4] DMPTool. URL: https://dmptool.org/ [Accessed: 22-Mar-2016]

[5] DMPOnline. URL: https://dmponline.dcc.ac.uk [Accessed: 22-Mar-2016]

[6] https://rd-alliance.org/groups/practical-policy-wg.html [Accessed: 22-Mar-2016]

[7] Chen, S. Y., Conway, M., Crabtree, J., Lee, C., Misra, S., Moore, R. W., . . . Xu, H. (2015). Policy Workbook -iRODS 4.0. Chapel Hill, North Carolina, USA: iRODS Consortium.

[8] Görzig, H., Engel, F., Brocks, H., Vogel, T., & Hemmje, H. (2015). Towards Data Management Planning Support for Research Data. iPRES2015. Chapel Hill, USA.

INDEX OF AUTHORS //

A
Kuldar Aas 57, 85, 311, 326
Mohammed Abo Ouda 270
Gabriel Abu-Tayeh 18
Matthew Addis 50, 276
Julie Allinson 50
Thorbjørn Ravn Andersen 260
Bethany Anderson 284
David Anderson 316
Chris Awre 50

B
Thomas Bähr 268
Mihai Bartha 85
Rob Baxter 330
Neil Beagrie 318
Frank Berghaus 168
Francine Berman 102
Henrike Berthold 54
Rhiannon Bettivia 212
Daniel Bicho 247
Aaron Birkland 80
Asger Askov Blekinge 260
Ashley Blewer 27
Jakob Blomer 168
Karl-Rainer Blumenthal 112
Eliane Blumer 306
Jose Borbinha 196
Cyril Bornet 186
Susan Braxton 284
Ena Brophy 272
Marcel Büchler 311, 326
Vincent Buntinx 186
Pierre-Yves Burgi 306

C
German Cancio Melia 168
Laura Capell 258
Bertrand Caron 32
Jonathan Clark 309, 320
Euan Cochrane 294
Louise Curham 266

D
Suenje Dallmeier Tiessen 168
Angela Dappert 113, 324
Kevin Davies 129
Michael Day 328
Tjarda de Haan 41
Kevin L. De Vorsey 328
Janet Delve 57, 85, 316, 326
Angela Di Iorio 32
Aude Dieudé 306
Tim Dilauro 80
Devan Ray Donaldson 95
Mark Donoghue 80
Heidi Dowding 95, 280, 304
Elise Dunham 284

E
Maud Ehrmann 186
Jürgen Enge 229
Felix Engel 330
Tom Ensom 139
Dragan Espenschied 37, 307

F
Patricia Falcao 139
Luis Faria 222, 311
Bruno Ferreira 222
Miguel Ferreira 222
Antonella Fresa 302
Yvonne Friese 268
Stefan Fritzsche 54

G
Gerado Ganis 168
Jay Gattuso 206, 328
Magnus Geber 302
Franziska Geisser 288
Urs Gerber 278
Thomas Gerdes 268
Helen Glaves 330
Andrea Goethals 290, 299, 328
Helen Gollin 278
Daniel Gomes 247
Heike Görzig 330
Richard Green 50
Peggy Griesinger 112
Myron Gutmann 102

H
Martin Häberle 149
Juha Hakala 309
Karen Hanson 80
Christen Hedegaard 72
Matthias Hemmje 330
Raquel Hill 95

I/J
Heidi Imker 284
Martin Iordanidis 300
Daniel Irrgang 149
Gregory Jansen 177
Claus Jensen 72
Joachim Jung 296
Bolette Ammitzbøll Jurik 260
Börje Justrell 302

K
Joanne Kaczmarek 77
Shuji Kamitsuna 286
Frédéric Kaplan 186
Roman Karl 85
Christian Keitel 95
William Kilbride 293, 296, 318
Julia Kim 112
Tsukasa Kimezawa 286
Steve Knight 299
Angelina Kraft 272
Heinz Werner Kramski 229
Ilya Kreymer 307
Thomas Kromann 237
Sabine Krug 54

L
Christopher Lee 256, 300, 316, 322
Thomas Liebetraut 37
Michelle Lindlar 66
Maurizio Lunghi 300, 309
Zoltan Lux 326

M
Jane Mandelbaum 299
Chiara Marciani 46
Richard Marciano 177
Alain Mast 46
Peter May 129
Nancy McGovern 299
Kenton McHenry 177
Peter McKinney 206, 313, 324
Evelyn McLellan 313, 324
Sam Meister 304
Kristof Meixner 159
Tobias Mérinat 46
Jessica Meyerson 294
Natasa Milic-Frayling 316, 294
David Minor 15
Jenny Mitcham 50
Jessica Moran 206
Sheila Morrissey 80
Sean Mosely 206
Michael Muraszko 280
Kate Murray 328

N
Anders Bo Nielsen 326
Caroline Nyvang 237

O
Krystyna Ohnesorge 46, 326
Achim Osswald 300

P
Frank Padberg 149, 282
Smruti Padhy 177
Dariusz Paradowski 264
Shira Peltzman 112, 304
Peter Pharow 302
Jan Potthoff 272
Claudio Prandoni 302
Stefan Proell 159
Diogo Proença 196

R
José Carlos Ramalho 222
Marcel Ras 23, 293, 296, 309, 320
Andreas Rauber 159
Matthias Razum 272
Klaus Rechert 37,139, 282
Carol Riccalton 320
Dave Rice 27
Kyle Rimkus 121, 284
Mike Ritter 15
Yannick Rochat 186
Andreas Romeyke 54
Jan Rörden 85
David Rosenthal 294

S
Jörg Sachse 54
Sibyl Schaefer 15, 299
Sven Schlarb 85
Martin Schlatter 278
Rainer Schmidt 85
Astrid Schoger 282
Andrea Ruth Schreiber 62
Sabine Schrimpf 293, 296
Natascha Schumann 300
Franziska Schwab 268
Ana Sesartic 306
Arif Shaon 270
Jamie Shiers 168
Barbara Sierman 293
Tibor Simko 168
Sarah Slade 262
Mike Smorul 15
Vicky Steeves 112
Tobias Steinke 282
Oleg Stobbe 37
Paul Stokes 318
Armin Straube 270
Matthias Stuermer 18
Roland Suri 288

T
Helen Tibbo 300
Philipp Tögel 149
Phillip Mike Tømmerholt 326
Matthias Töwe 288
Gail Truman 290, 299

V

Remco Van Veenendaal 309
Erwin Verbruggen 302
Ricardo Vieira 196
Zach Vowell 294

W

Simon Waddington 330
Jaye Weatherburn 274
Brent West 77
Paul Wheatley 328
David Wilcox 315
Elliot Williams 258
Andrew Wilson 57
Carl Wilson 304
Simon Wilson 50
Scott Witmer 121
Andrew Woods 315
Kam Woods 256

Y

Jeremy York 102

Z

Eld Zierau 237, 313, 324
Kate Zwaard 299

iPRES 2016

Bern // October 3 – 6, 2016

Proceedings of the 13th International Conference on Digital Preservation