iPRES 2015

CHAPEL HILL – NOVEMBER 2–6

Proceedings of the 12th International Conference on Digital Preservation

Proceedings of the 12th International Conference on Digital Preservation

ISBN 10-0-692-59881-2

ISBN 13-978-0-692-59881-8

School of Information and Library Science
The University of North Carolina at Chapel Hill
216 Lenoir Drive, CB #3360, 100 Manning Hall
Chapel Hill, NC 27599-3360
info@ils.unc.edu

Hosted by

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

UNC UNIVERSITY LIBRARY

UNC THE ODUM INSTITUTE

Sponsored by

DPN THE DIGITAL PRESERVATION NETWORK

ExLibris The bridge to knowledge

ORACLE®

libnova DIGITAL PRESERVATION SOFTWARE

Preservica Digital Preservation

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE

UNC UNIVERSITY LIBRARY

UNC THE ODUM INSTITUTE

iPRES 2015

CHAPEL HILL – NOVEMBER 2–6

Proceedings of the 12th International Conference on Digital Preservation

Table of Contents

Organizing Committee

Jonathan Crabtree	Odum Institute for Research in Social Science (Posters and Demos Co-Chair)
Leo Konstantelos	University of Melbourne (Program Co-Chair)
Christopher (Cal) Lee	University of North Carolina (General Co-Chair)
Nancy McGovern	Massachusetts Institute of Technology Libraries (Program Co-Chair)
Yukio Maeda	University of Tokyo (Posters and Demos Co-Chair)
Maureen Pennock	British Library (Workshops and Tutorials Co-Chair)
Helen Tibbo	University of North Carolina (General Co-Chair)
Kam Woods	University of North Carolina (Digital Preservation Showcase Chair)
Eld Zierau	Royal Library of Denmark (Workshops and Tutorials Co-Chair)

Program Committee

Thu-Mai Christian	University of North Carolina [US]
Sandra Collins	Digital Repository Ireland (Dublin) [Ireland]
Libor Coufal	National Library of Australia [Australia]
Jonathan Crabtree	Odum Institute for Research in Social Science [US] (Posters and Demos Co-Chair)
Janet Delve	University of Portsmouth [UK]
Milena Dobreva	University of Malta [Malta]
Andrea Goethals	Harvard [US]
Neil Grindley	JISC [UK]
Carolyn Hank	University of Tennessee [US]
Christy Henshaw	Wellcome Trust [UK]
Leslie Johnston	National Archives and Records Administration [US]
Catherine Jones	Science & Technology Facilities Council [UK]
Mark Jordan	Simon Fraser University [Canada]
Ulla Bøgvad Kejser	Royal Library of Denmark [Denmark]
Leo Konstantelos	University of Melbourne [Australia] (Program Co-Chair)
Chris Lacinak	AVPreserve [US]
Christopher (Cal) Lee	University of North Carolina [US] (General Co-Chair)
Michelle Lindlar	German National Library of Science and Technology [Germany]
Gavan McCarthy	University of Melbourne [Australia]
Nancy McGovern	Massachusetts Institute of Technology Libraries [US] (Program Co-Chair)
Yukio Maeda	University of Tokyo [Japan] (Posters and Demos Co-Chair)
Steve Marks	University of Toronto [Canada]
Jessica Moran	National Library of New Zealand [New Zealand]
Courtney Mumma	Artefactual Systems [US]
Kate Murray	Library of Congress [US]
Dave Pcolar	Digital Preservation Network [US]
Maureen Pennock	British Library [UK] (Workshops and Tutorials Co-Chair)
Klaus Rechert	University of Freiburg [Germany]
Gabby Redwine	Yale University [US]
Richard Rinehart	Samek Art Museum, Bucknell University [US]
João Rocha da Silva	University of Porto [Portugal]
Daisy Selematsela	National Research Foundation [South Africa]
Katherine Skinner	Educopia Institute [US]
Armin Straube	Nestor [Germany]
Shigeo Sugimoto	University of Tsukuba [Japan]
Manfred Thaller	University at Cologne [Germany]
Susan Thomas	Oxford University [UK]
Helen Tibbo	University of North Carolina [US] (General Co-Chair)
Lucia Maria	Velloso de Oliveira, Fundacao Casa de Rui Barbosa [Brazil]
Kam Woods	University of North Carolina [US] (Digital Preservation Showcase Chair)
Eld Zierau	Royal Library of Denmark [Denmark] (Workshops and Tutorials Co-Chair)
Kate Zwaard	Library of Congress [US]

Preface

Co-conveners Christopher (Cal) Lee and Helen Tibbo of the University of North Carolina at Chapel Hill welcomed 327 delegates from 22 countries to Chapel Hill for the 12th International Conference on Digital Preservation (iPres), held November 2-6, 2015.

The conference was structured around a program of workshops and tutorials on Monday and Friday, with papers, posters and panels during the core conference from Tuesday to Thursday. In response to feedback from delegates in previous years, the organizers experimented with a set of interactive sessions that included an opening session with Twitter participation that highlighted digital preservation developments since the 2014 iPres conference; a closing session noting projects and initiatives to watch in the lead up to the 2016 iPres conference; a facilitated community discussion on preservation storage; a policy and documentation clinic; and create-your-own meeting time slots called "Get a Room" sessions to allow groups to convene on any topic of interest.

We received 110 total submissions, and the final program included 12 long papers, 15 short papers and 33 posters, 3 demos, 6 workshops, 3 tutorials and 5 panels. To ensure a broad representation of innovative work on digital preservation while also maintaining high standards for the text of papers, we included a "revise and resubmit" category for submissions; and we were happy that all authors of papers who were extended the opportunity to revise their papers for reconsideration did so, resulting in much stronger contributions in all cases. As a result, we were able to ultimately include 60% (27 out of 45) of the submitted papers in these proceedings.

Keynotes

Lisa Nakamura is the Gwendolyn Calvert Baker Collegiate Professor in the Department of American Culture and the Department of Screen Arts and Cultures at the University of Michigan, Ann Arbor, where she also serves as Coordinator of Digital Studies. She has written extensively on issues of race, gender, and sexuality in digital media. She co-facilitates FemTechNet, an active network of artists, researchers, activists, students, and librarians engaged at the intersections of science, feminism, and technology. She presented a case study called, "The Digital Afterlives of This Bridge Called My Back: Public Feminism and Open Access" that concluded with lessons learned for digital preservation.

Pamela Samuelson is the Richard M. Sherman Distinguished Professor of Law and Information at the University of California, Berkeley with a joint appointment in the UC Berkeley School of Information and School of Law. She has been a pioneer in issues of cyberlaw, intellectual property rights, and digital copyright law as well as an active and influential voice in critical discussions of information policy. Her presentation "Mass Digitization of Cultural Heritage: Can Copyright Obstacles Be Overcome?" drew upon her extensive experience as applied to digital collections and preservation issues.

The program

The conference placed an emphasis on research and innovative practice while focusing on core themes in digital preservation: Frameworks for Digital Preservation; Institutional Opportunities and Challenges; Infrastructure Opportunities and Challenges; Preservation Strategies and Workflows; and Digital Preservation Frameworks, Strategies and Workflows; and Dimensions of Digital Preservation. The program committee received an excellent batch of papers and posters that examined a range of timely topics, including digital preservation policies, web archiving, digital art, digital forensics, preservation storage, collaborative preservation planning, preservation metadata, research data management, and preservation costs.

The award for Best Paper (sponsored by Ex Libris) went to Reagan Moore, Arcot Rajasekar and Hao Xu for their paper "DataNet Federation Consortium Preservation Policy Toolkit." The judges noted that a great strength of the paper was its clear roots in practical scenarios, developing a practical and usable set of policies, specifically within the framework of iRODS but clearly transferable to other technical environments. They particularly liked the illustration of how it can be used with ISO16363, thereby complementing current tools. Honorable mentions went to Stephen Abrams for "A Foundational Framework for Digital Curation: The Sept Domain Model" and Douglas

Thain, Peter Ivie and Haiyan Men for "Techniques for Preserving Scientific Software Executions: Preserve the Mess or Encourage Cleanliness?"

The award for Best Poster (sponsored by School of Information and Library Science at UNC) went to Bolette Ammitzbøll Jurik and Asger Askov-Blekinge for their poster on "Minimal Effort Ingest."

Last year's Digital Preservation Showcase was a great success, so we repeated it this year with a twist: focusing more on support for digital preservation decision making. Presenters representing Archivematica, Islandora, Preservica, and Rosetta used descriptions, examples, and demonstrations to address a set of questions on Ingest, Preservation and Storage, and Access with an open discussion as the wrap-up.

Acknowledgments

This year's conference was generously supported by the following sponsors: Digital Preservation Network (DPN), Ex Libris, Libnova, Oracle, Preservica, the Odum Institute for Research in Social Science at the University of North Carolina at Chapel Hill (UNC-CH), the UNC-CH Libraries and the UNC-CH School of Information and Library Science (SILS).

The Organizing Committee was delighted with the success of the conference and acknowledges with gratitude the contributions of the many Program Committee members who worked so hard to ensure that the program included papers, posters, demonstrations, workshops, and tutorials of the highest quality. The warm hospitality extended to delegates by local organizers throughout the week made the conference both productive and enjoyable.

Finally, we would like to thank staff from the Odum Institute and the numerous student volunteers who worked hard behind-the-scenes work to ensure that the conference ran so smoothly. Preparations for next year are underway by the Swiss National Library, host for iPres 2016. We hope to see you all there.

Organizing Committee, iPres 2015

iPRES 2015 Program

Monday, November 2, 2015
Workshops & Tutorials

8am-5pm **REGISTRATION**

9am-3pm **Fedora 4 Tutorial**

Andrew Woods (Fedora 4 Technical Lead)

This tutorial will provide an introduction to and overview of Fedora 4, with a focus on the latest features. Fedora 4 implements the W3C Linked Data Platform recommendation, so a section of the tutorial will be dedicated to a discussion about LDP and the implications for Fedora 4 and linked data. Fedora 4 is also designed to be integrated with other applications, so a section of the tutorial will review common applications and integration patterns. Finally, attendees will participate in a hands-on session that will give them a chance to install, configure, and explore Fedora 4 by following step-by-step instructions.

9am-5pm **Testing the Proposed METS 2.0 Data Model**

Bertrand Caron (Department of Bibliographic and Digital Information, Bibliothèque nationale de France), Andreas Nef (Docuteam GmbH), Thomas G. Habing (Library Software Development Group, University of Illinois at Urbana-Champaign) and Nancy J. Hoebelheinrich (Knowledge Motifs LLC, San Mateo, CA , USA)

In this workshop, participants will first develop an understanding of the data models underlying some canonical uses of the existing METS schema as a contextual basis for the description of a next generation METS (2.0) data model. Following the description of the METS 2.0 data model, a number of use cases applying the proposed data model will be discussed to address questions such as how the METS 2.0 data model fits existing implementations, what issues arise from that application, and whether there are more opportunities than challenges to the evolution of the data model as currently proposed. Finally, to put the proposed METS data model into a broader context, complementary data models currently being developed will be discussed such as SEDA (Data Exchange Standard for Archiving) or the Portland Common Data Model. Participants will be invited to participate in the discussions, and the evaluation / refinement of a METS 2.0 data model.

Roles & Responsibilities for Sustaining Open Source Platforms & Tools

Trevor Owens (Senior Program Officer, Institute of Museum and Library Services)

This workshop invites stakeholders representing organizations that play different roles in the open source software ecosystem to share their respective perspectives on roles in this ecosystem. Through discussion, participants will work through issues as they relate to different kinds of open source software systems. These include: 1) descriptions of roles that should be in play as open source software projects move from research/startup phases toward implementation, dissemination, and ultimately maintenance and ongoing feature development; 2) the role of project-based funding; 3) the tradeoffs around different open source software sustainability models; and 4) the role that education, training and ongoing professional development plays in ensuring the use and maintenance of these tools and platforms.

From Theory to Practice: Using ISO16363

Helen R. Tibbo (UNC – Chapel Hill), Nancy Y. McGovern (MIT Libraries), Barbara Sierman (National Library of the Netherlands), Ingrid Dillo (DANS: Data Archiving & Networked Services) and Courtney Mumma (Artefactual Systems, Inc.)

The ISO16363 Standard is a formal framework for determining whether an organization is a Trustworthy Digital Repository. Published in 2012, the standard considers not only the technical infrastructure used for digital object management but also organizational infrastructure, and security risk management. Recognizing that this can go beyond the experience of many new users. This tutorial will focus on an array of options and programs for audit and potential certification of trustworthy digital repositories. These will include self-audit, the European three-level model of certification, the Data Seal of Approval, peer-audit, ISO 16363 audit, and forthcoming certification of trustworthy repositories.

6:00pm-8:00pm **OPENING RECEPTION**
Venue: Wilson Library, University of North Carolina at Chapel Hill
Sponsored by Oracle
Entertainment provided by Clef Hangers

The UNC Clef Hangers (also known as the Clefs) is the oldest a cappella group at the University of North Carolina at Chapel Hill. The Clef Hangers were established in 1977, and were originally called the Morrison Dorm Singers. In their first concert in 1979, they wore vests (covered with buttons) and bowties, which they continue to don today. Since their first tour to New Orleans, Louisiana in 1980, the Clefs have performed for audiences in Spain, Mexico, France, Scotland, Switzerland, The Bahamas, Los Angeles, Washington DC, New York, and many other locations domestic and abroad. During a tour to New York City, they performed on the television show Good Morning America. Since 2002, they have also performed at the UNC Commencement ceremony. The Clefs have released several professionally produced studio albums, which have received numerous awards.

Transportation will be provided from the Friday Center to the venue.

Tuesday, November 3, 2015

8am-5pm **REGISTRATION**

9am-10am **Opening and Welcome:**
Helen Tibbo (General Co-Chair), Alumni Distinguished Professor, School of Information and Library Science, UNC-Chapel Hill
Christopher (Cal) Lee (General Co-Chair), Associate Professor, School of Information and Library Science, UNC-Chapel Hill
Sally Greene, Chapel Hill Town Council member, Mayor Pro Tempore
Sarah Michalak, Associate Vice Provost and University Librarian, UNC-Chapel Hill
Tom Carsey, Thomas J. Pearsall Distinguished Professor, Department of Political Science, and Director of the Odum Institute for Research in Social Science, UNC-Chapel Hill

Spotlight: This Year's Digital Preservation Noteworthy Progress and Achievements
Facilitator: Nancy McGovern, Head, Curation and Preservation Services, MIT Libraries

"What would you highlight as significant developments or outcomes over the past year?" The facilitator and presenters will use this question as the basis for a brief review of digital preservation highlights for the year. A couple examples to illustrate: PREMIS version 3.0 was announced and the next OAIS Reference Model revision is underway. Other examples from the digital preservation community might point to: project results; revisions of standards, tools, or software released; or indicators of program developments (e.g., policies developed and shared, preservation strategies demonstrated). In addition to examples from the facilitators to get things going, a core of the review will build on themes and examples from attendee contributions and tweets from across the digital preservation community in response to this question - please join in! What would you highlight?

10am-10:30am **BREAK**

10:30am-12pm **SESSIONS**

Frameworks for Digital Preservation *Grumman Auditorium*

One Core Preservation System For All Your Data. No Exceptions! (L)* Marco Klindt and Kilian Amrhein

In this paper, we describe an OAIS aligned data model and architectural design that enables us to archive digital information with a single core preservation workflow. The data model allows for normalization of metadata from widely varied domains to ingest and manage the submitted information utilizing only one generalized toolchain and be able to create access platforms that are tailored to designated data consumer communities. The design of the preservation system is not dependent on its components to continue to exist over its lifetime, as we anticipate changes both of technology and environment. The initial implementation depends mainly on the open-source tools Archivematica, Fedora/Islandora, and iRODS.

10:30am-12pm **A Foundational Framework for Digital Curation: The Sept Domain Model.** (L) Stephen Abrams

Digital curation is a complex of actors, policies, practices, and technologies enabling successful consumer engagement with authentic content of interest across space and time. While digital curation is a rapidly maturing field, it still lacks a convincing unified theoretical foundation. A recent internal evaluation of its programmatic activities by the University of California Curation Center (UC3) led quickly to seemingly simple, yet deceptively difficult-to-answer questions. Too many fundamental terms of curation practice remain overloaded and under-formalized, perhaps none more so than "digital object." To address these concerns, UC3 is developing a new model for conceptualizing the curation domain. While drawing freely from many significant prior efforts (e.g., Kahn-Wilensky, FRBR, NAA, OAIS, BRM, etc.), the UC3 Sept model also assumes that digital curation is an inherently semiotic activity. Consequently, the model considers curated content with respect to six distinct analytic dimensions: semantics, syntactics, empirics, pragmatics, diplomatics, and dynamics, which refer respectively to content's underlying abstract meaning or emotional affect, symbolic encoding structures, physical representations, realizing behaviors, evidential authenticity and reliability, and evolution through time. Correspondingly, the model defines an object typology of increasing consumer utility: blobs, artifacts, exemplars, products, assets, records, and heirlooms, which are respectively existential, intentional, purposeful, interpretable, useful, trustworthy, and resilient digital objects. Content engagement is modeled in terms of producer, owner, manager, and consumer roles acting within a continuum of concerns for originating, organizing, and pluralizing curated content. Content policy and strategy are modeled in terms of six high-level imperatives: predilect, collect, protect, introspect, project, and connect. A consistent, comprehensive, and conceptually parsimonious domain model is important for planning, performing, and evaluating programmatic activities in a rigorous and systematic rather than ad hoc and idiosyncratic manner. The UC3 Sept model can be used to make precise yet concise statements regarding curation intentions, activities, and results.

Developing a Framework for File Format Migrations. (L) Joey Heinen and Andrea Goethals.

In this paper, we describe the development of a file format migrations framework at Harvard Library, using one migration case study, Kodak PhotoCD images, to demonstrate implementation of the framework.

Institutional Opportunities and Challenges

Benchmarks for Digital Preservation Tools. (L) Kresimir Duretec, Artur Kulmukhametov, Andreas Rauber and Christoph Becker

Creation and improvement of tools for digital preservation is a difficult task without an established way to assess any progress in their quality. This happens due to low presence of solid evidence and a lack of accessible approaches to create such evidence. Software benchmarking, as an empirical method, is used in various fields to provide objective evidence about the quality of software tools. However, digital preservation field is still missing a proper adoption of that method. This paper establishes a theory of benchmarking of tools in digital preservation as a solid method for gathering and sharing the evidence needed to achieve widespread improvements in tool quality. To this end, we discuss and synthesize literature and experience on the theory and practice of benchmarking as a method and define a conceptual framework for benchmarks in digital preservation. Four benchmarks that address different digital preservation scenarios are presented. We compare existing reports on tool evaluation and how they address the main components of benchmarking, and we discuss the question of whether the field possesses the right combination of social factors that make benchmarking a promising method at this point in time. The conclusions point to significant opportunities for collaborative benchmarks and systematic evidence sharing, but also several major challenges ahead.

Towards a Common Approach for Access to Digital Archival Records in Europe. (L) Alex Thirifays and Kathrine Hougaard Edsen Johansen

This paper describes how the E-ARK project (European Archival Records and Knowledge Preservation) aims to develop an overarching methodology for curating digital assets. This methodology must address business needs and operational issues, proposing a technical wall-to-wall reference implementation for the core OAIS flow – Ingest, Archival Storage and Access. The focal point of the paper is the Access part of the OAIS flow. The paper first lays out the access vision of the E-ARK project, and secondly describes the method employed to enable information processing and to pin-point the functional and non-functional requirements. These requirements will allow the E-ARK project to create a standardized format for the Dissemination Information Package (DIP), and to develop the access tools that will process this format. The paper then proceeds to describe the actual DIP format before detailing what the access solution will look like, which tools will be developed and, not least, why the E-ARK Access system will be used and work.

10:30am-12pm **Developing a Highly Automated Web Archive System Based on IIPC Open Source Software.** (S) Zhenxin Wu, Jin Xie, Jiying Hu and Zhixiong Zhang

In this paper, we describe our development of a highly automated web archiving system based on IIPC open source software at the National Science Library (NSL). We designed a web archiving platform which integrates with popular IIPC tools, as well as developing several modules to meet special requirements of the NSL. We have applied a cooperative mode of central management server and collecting client, which can complete the unified management of seeds and support the collaborative work of multiple crawlers. Some modules were developed to improve the automation of web archiving workflow and provide more services.

Best Until … A National Infrastructure for Digital Preservation in the Netherlands. (S) Barbara Sierman and Marcel Ras

This paper describes the developments in the Netherlands to establish a national Network for Digital Heritage. This network is based on three pillars: to make the digital heritage visible, usable and sustainably preserved. Three working programmes will have 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.

Panel

Good, Better, Best? Examining the Range and Rationales of Institutional Data Curation Practices. Robin Rice, Limor Peer, Wendy White and Florio Arguillas

Many academic institutions are grappling with managing local research data assets. Resources and approaches vary. This panel will explore curation procedures at institutional data repositories.

12pm-1pm **LUNCH**

1pm-2:30pm **SESSIONS**

Infrastructure Opportunities and Challenges

Archiving Deferred Representations Using a Two-Tiered Crawling Approach. (L) Justin Brunelle, Michele Weigle and Michael Nelson

Web resources are increasingly interactive, resulting in resources that are increasingly difficult to archive. The archival difficulty is based on the use of client-side technologies (e.g., JavaScript) to change the client-side state of a representation after it has initially loaded. We refer to these representations as deferred representations. We can better archive deferred representations using tools like headless browsing clients. We use 10,000 seed Universal Resource Identifiers (URIs) to explore the impact of including PhantomJS – a headless browsing tool – into the crawling process by comparing the performance of wget (the baseline), PhantomJS, and Heritrix. Heritrix crawled 2.065 URIs per second, 12.15 times faster than PhantomJS and 2.4 times faster than wget. However, PhantomJS discovered 531,484 URIs, 1.75 times more than Heritrix and 4.11 times more than wget. To take advantage of the performance benefits of Heritrix and the URI discovery of PhantomJS, we recommend a tiered crawling strategy in which a classifier predicts whether a representation will be deferred or not, and only resources with deferred representations are crawled with PhantomJS while resources without deferred representations are crawled with Heritrix. We show that this approach is 5.2 times faster than using only PhantomJS and creates a frontier (set of URIs to be crawled) 1.8 times larger than using only Heritrix.

Techniques for Preserving Scientific Software Executions: Preserve the Mess or Encourage Cleanliness? (L) Douglas Thain, Peter Ivie and Haiyan Meng

An increasing amount of scientific work is performed in silico, such that the entire process of investigation, from experiment to publication, is performed by computer. Unfortunately, this has made the problem of scientific reproducibility even harder, due to the complexity and imprecision of specifying and recreating the computing environments needed to run a given piece of software. Here, we consider from a high level what techniques and technologies must be put in place to allow for the accurate preservation of the execution of software. We assume that there exists a suitable digital archive for storing digital objects; what is missing are frameworks for precisely specifying, assembling, and executing software with all of its dependencies. We discuss the fundamental problems of managing implicit dependencies and outline two broad approaches: preserving the mess, and encouraging cleanliness. We introduce three prototype tools for preserving software executions: Parrot, Umbrella, and Prune.

1pm-2:30pm **A Method for the Systematic Generation of Audit Logs in a Digital Preservation Environment and Its Experimental Implementation In a Production Ready System.** (S) Hao Xu, Jason Coposky, Dan Bedard, Jewel Ward, Terrell Russell, Arcot Rajasekar, Reagan Moore, Ben Keller and Zoey Greer

In a digital preservation environment there is a need for a complete auditing of system state changes. A complete log ensures that the properties of the objects in the system can be verified. Modern data management systems such as the integrated Rule-Oriented Data System (iRODS) allow administrators to configure complex policies. Pre- or post-operation, these policies can trigger other state changing operations. In this paper, we describe a method that allows us - given a complete list of state changing operations - to generate a complete audit log of the system. We also describe an experimental implementation of the framework. An important advantage of our method is that not only do we build on sound theoretical foundations, but we also validate the methodology in a production-ready environment which has undergone substantial quality control. The implementation of our method can be distributed as a turnkey solution that is ready to deploy, which significantly shortens the gap between theoretical development and practical applications.

Preservation Strategies and Workflows

Functional Access to Forensic Disk Images in a Web Service. (S) Kam Woods, Christopher Lee, Oleg Stobbe, Thomas Liebetraut and Klaus Rechert

We describe a hybrid approach for access to digital objects contained within forensic disk images extracted from physical media. This approach includes the use of emulation-as-a-service (EaaS) to provide web-accessible virtual environments for materials that may not render or execute accuratelyon modern hardware and software, and the use of digital forensics software libraries to produce web-accessible file system views to support single-file access and provide visualizations of the file system.

Experiment, Document & Decide: A Collaborative Approach to Preservation Planning at the BnF. (S) Bertrand Caron, Thomas Ledoux, Jean-Philippe Tramoni and Stéphane Reecht

The National Library of France (BnF) has recently implemented a new module for its Scalable Preservation and Archiving Repository (SPAR) to set up preservation strategies based on formats, agents, workflows, tools and tests, and managed as reference packages in the Archive. This module aims to fulfill an objective: for SPAR to be fully self-documented. Formats, agents and workflows are formally described and preserved along with the Information packages in which such elements are involved. Although this was a feature that was included from the beginnings of SPAR, the new Preservation Planning module aims to provide a tool that can more easily build these reference packages and that will more closely involve domain experts and the IT department in the processes of preservation planning. But the main innovation lies in the documentation of decisions that directed their selection as standards in SPAR: test data are now preserved as a new kind of reference package.

Beyond the Binary: Pre-Ingest Preservation of Metadata. (S) Jessica Moran and Jay Gattuso

This paper describes some of the challenges the National Library of New Zealand has faced in our efforts to maintain the authenticity of born digital collection items from first transfer to the Library through ingest into our digital preservation system. We assume that assuring the authenticity and integrity of digital objects means preserving the binary objects plus metadata about the objects. We discuss the efforts and challenges of the Library to preserve contextual metadata around the binary object, in particular filenames and file dates. We discuss these efforts from the two perspectives of the digital archivist and the digital preservation analyst, and how these two perspectives inform our current thinking.

Preserving an Evolving Collection: "On-The-Fly" Solutions for the Chora of Metaponto Publication Series. (S) Jessica Trelogan, Maria Esteva and Lauren Jackson

As digital scholarship continues to transform research, so it changes the way we present and publish it. In archaeology, this has meant a transition from the traditional print monograph, representing the "definitive" interpretation of a site or landscape, to an online, open, and interactive model in which data collections have become central. Online representations of archaeological research must achieve transparency, exposing the connections between fieldwork and research methods, data objects, metadata, and derived conclusions. Accomplishing this often requires multiple platforms that can be burdensome to integrate and preserve. To address this, the Institute of Classical Archaeology and the Texas Advanced Computing Center have developed a "collection architecture" that integrates disparate and distributed cyberinfrastructure resources through a customized automated metadata platform, along with procedures for data presentation and preservation. The system supports "on-the-fly" data archiving and publication, as the collection is organized, shared, documented, analyzed, and distributed.

1pm-2:30pm **Participatory Digital Repositories for the Curation of Performing Arts with Digital Technology.** (S) Guillaume Boutard

The complexity of socio-technical systems in artistic production involving digital technology, especially in the performing arts, challenges digital curation models with a potential shift from cycles to networks. We argue that digital curation models need to develop in parallel to interdisciplinary investigations of these systems. These investigations question the conceptual separation of curation stages as well as roles. In this paper, we build on previous curation projects for new media arts and on the historical analysis of a specific work of contemporary music with live electronics to propose future directions for the integration of curation practices, artistic practices and digital curation models.

Characterization of CDROMs for Emulation-based Access. (L) Klaus Rechert, Thomas Liebetraut, Oleg Stobbe, Isgandar Valizada and Tobias Steinke

Memory institutions have already collected a large number of digital objects, predominantly CD-ROMs. Some of them are already inaccessible with current systems, and most of them will be soon. Emulation offers a viable strategy for long-term access to these publications. However, these collections are huge and the objects are missing technical metadata to setup a suitable emulated environment. In this paper we propose a pragmatic approach to technical metadata which we use to implement a characterization tool to suggest a suitable emulated rendering environment.

Panel

Advancing the Evidence Base of Digital Preservation. Micah Altman, Helen Tibbo and NDSA Coordinating Committee

Research is critical to the advancement of both a basic understanding and the effective practice of digital preservation. Research must, however, be intimately linked to practice in order to improve outcomes. This panel will discuss methodology, metrics, tools, and exemplars that can effectively build the evidence base for digital preservation. Panelists will present on a simulation framework for evaluating preservation risks, formal/machine actionable preservation strategies and implementations; and evaluation of preservation performance.

2:30pm-3pm **BREAK**

3pm-4:30pm **SESSIONS**

Institutional Opportunities and Challenges

DataNet Federation Consortium Preservation Policy Toolkit. (L) Reagan Moore, Arcot Rajasekar and Hao Xu

The DataNet Federation Consortium uses a policy-based data management system to apply and enforce preservation requirements. This paper describes the Preservation Policy Toolkit developed by the consortium. In particular, the paper describes the infrastructure needed for preservation, presents examples of computer actionable forms of policies, and provides a generic template for designing actionable preservation policies.

Preserving the Fruit of Our Labor: Establishing Digital Preservation Policies and Strategies at the University of Houston Libraries. (S)Santi Thompson, Annie Wu, Drew Krewer, Mary Manning and Rob Spragg

To develop a comprehensive digital preservation program for maintaining long-term access to the Libraries' digital assets and align our practices with national standards and guidelines, the University of Houston (UH) Libraries formed the Digital Preservation Task Force (DPTF) to assess previous digital preservation practices and make recommendations on future efforts. This paper outlines the methodology used, including the task force's use of existing models and evaluation criteria, to successfully generate new policies and select Archivematica as our system to process and preserve our digital assets. It concludes with recommended strategies for the implementation of the policies and preservation operations.

Copyright and the Digitization of State Government Documents: A Preliminary Analysis. (S) Brett Currier, Anne Gilliland and David Hansen

In this paper we explore the copyright status of state and local government documents and address some of the legal issues encountered when digitizing them.

3pm-4:30pm **Project Chrysalis – Transforming the Digital Business of the National Archives of Australia.** (S) Zoe D'Arcy

The role of the National Archives of Australia is to promote the creation, management and preservation of authentic, reliable and usable Commonwealth government records and enable ongoing public access to the archival resources of the Commonwealth. Records that are created by Commonwealth government agencies and transferred to the National Archives are, of course, predominately digital. Digital records bring a range of challenges, but they also potentially present new opportunities in the way archives can conduct their business. This paper outlines a project currently underway at the National Archives, named Project Chrysalis, which is an end-to-end business system that aims to transform the way in which the Archives does its digital business. Project Chrysalis represents not just a technical solution, but also significant business change for the National Archives. However, if implemented successfully, the project should enable the Archives to sustainably harvest, preserve and provide access to digital records in the information age.

Digital Preservation Frameworks, Strategies and Workflows

Lessons Learned and Open Challenges Regarding Data Management Plans and Research Date Management Support (L) Heike Görzig, Felix Engel, Matthias L. Hemmje and Holger Brocks

This paper outlines an approach for developing tools and services that support automated generation, management, evolution and execution of data management plans (DMPs) by generating rules derived from the DMPs which can be applied to the data to be archived. The approach is based on existing models and tools that were developed in successive research projects SHAMAN, APARSEN, and SCIDIP-ES. The models include the Curation Lifecycle Model from the DCC, the OAIS Information Model and the Extended Information Model to support processes, domains, and organizations. An approach for deriving rules from policies is outlined to support using iRODS. OAIS and Context Information related to a data object is supported in a serialization using the OAI-ORE format.

Human and Machine-Based File Format Endangerment Notification and Recommender Systems Development. (S) Heather Ryan, Roman Graf and Sergiu Gordea

Effectively preserving access to digital content over time is dependent on availability of an appropriate IT infrastructure including access to appropriate rendering software and its requisite operating systems and hardware. The complexity of this task increases over time and with the size and heterogeneity of digital collections. Automating notifications on file format endangerment and decision recommendations can greatly improve preservation planning processes. This paper presents work in progress that contributes to the design and testing of an automated file format endangerment notification and recommendation system. This system's design is based on concepts explored in previous research, but it presents the novel application of statistically generated similarity profiles and machine-generated recommendations based on human expert input.

Deduplicating Bibliotheca Alexandrina's Web Archive. (S) Youssef Eldakar and Magdy Nagi

Archiving web content is bound to produce datasets with duplication, either across time or across location. The Bibliotheca Alexandrina (BA) has a web archive legacy spanning a period of 10 years and is continuing to expand the collection. Initial assessment of this very large store of data was conducted. Given a high enough rate of duplication, deduplication would lead to sizable savings in storage requirements. The BA worked through the International Internet Preservation Consortium (IIPC) to compile best practices for recording duplicates in ISO 28500, the WARC File Format. To deduplicate legacy web archives "after the fact," the BA is implementing the WARCrefs deduplication tools. Following implementation and testing, the BA plans to put the tools to use to deduplicate its one petabyte of archived web content.

Panel

Preserving Born-Digital News. Edward McCain, Hannah Sommers, Christie Moffatt, Abigail Potter, Stéphane Reecht and Martin Klein

The news industry has quickly adopted networked digital technologies to create and distribute their content across all media types and in an ever-increasing number of formats. These technologies have also enabled individuals to capture and share information, news, and opinion on contemporary and community events. These changes contribute to a dynamic news ecosystem, upending traditional publishing models that media companies, libraries, archives and memory institutions have depended on to save the news. In this panel, the challenges and opportunities of preserving born-digital news content will be presented and discussed. A preliminary environmental scan of the state of digital news preservation will be shared. Perspectives and tactics from the "front-line" of news creation will be covered in addition to establishing special collections to capture and preserve web sites that cover news events. Efforts to establish relationships with the creators of content management systems (CMS) that drive the back end of modern media publishing networks will also be presented, as will tools that have been developed to capture social media and other content from the web that contributes to the present day news ecosystem.

6:30pm-8:30pm **GROUP DINNER**

Venue: Carolina Club, George Hill Watts Alumni Center

Performance by The Bluegrass Experience

Tommy Edwards is lead singer and guitarist for The Bluegrass Experience, one of the Southeast's most respected traditional music groups. He is also host of the "Bluegrass Saturday Night" radio program, which features both classic and contemporary bluegrass recordings as well as interviews with people associated with the music and a calendar of traditional music events in the Heart of Carolina. Edwards has performed professionally for more than 35 years, was twice named World Champion Bluegrass Guitarist and has recorded or performed live with an array of bluegrass greats. He and his wife Cindy are respected collectors of and authorities on the traditional pottery of central North Carolina. They operated an antiques business for more than 25 years.

Transportation will be provided from the Friday Center to hotels. The buses will then run to and from hotels (except Carolina Inn) to the venue.

Wednesday, November 4, 2015

9am-10am **KEYNOTE**

The Digital Afterlives of This Bridge Called My Back: Public Feminism and Open Access
Lisa Nakamura, Gwendolyn Calvert Baker Collegiate Professor, Department of American Cultures and the Department of Screen Arts and Cultures, University of Michigan

Chair: Lilly U. Nguyen, Assistant Professor, Department of Women's and Gender Studies, University of North Carolina at Chapel Hill

This presentation will describe how the social media platform Tumblr has been deployed by fans as a site of memory for the canonical and until recently out of print woman of color text This Bridge Called My Back. The curation, distribution, and communities of shared feeling that have formed around this text demonstrate how it has come to function as a rallying point for post-digital feminists.

10am-10:30am **BREAK**

10:30am-12pm **SESSIONS**

Dimensions of Digital Preservation

Applying Translational Principles to Data Science Curriculum Development. (L) Liz Lyon, Eleanor Mattern, Amelia Acker and Alison Langmead

This paper reports on a curriculum mapping study that examined job descriptions and advertisements for three data curation focused positions: Data Librarian, Data Steward / Curator, and Data Archivist. We present a transferable methodological approach for curriculum development and the findings from our evaluation of employer requirements for these positions. This paper presents "model pathways" for these data curation roles and reflects on opportunities for iSchools to adopt translational data science principles to frame and extend their curriculum to prepare their students for data-driven career opportunities.

Educational Records of Practice: Preservation and Access Concerns. (S) Elizabeth Yakel, Rebecca Frank and Kara Suzuka

Researchers in information science are placing increased attention on data reuse and on what must be preserved with that data to enable meaningful use by scholars within and across disciplines. Although the focus has been on scientific or quantitative data, this paper expands the discussion to qualitative data – specifically digital video records of practice in the field of education. This is an interesting case because researchers and diverse education professionals are interested in reusing this content, though their needs differ. We focus on three issues that raise challenges for preservation and access: file format, context, and dissemination.

10:30am-12pm **A Survey of Organizational Assessment Frameworks in Digital Preservation.** (L) Emily Maemura, Nathan Moles and Christoph Becker

As the field of digital preservation continues to mature, there is an increasing need to systematically assess an organization's abilities to achieve its digital preservation goals. A wide variety of assessment tools exist for this purpose. These range from light-weight checklists to resource-intensive certification processes. Conducted as part of the BenchmarkDP project, this paper presents a survey of these tools that elucidates available options for practitioners and opportunities for further research.

Getting to the Bottom Line: 20 Digital Preservation Cost Questions. (S) Matt Schultz, Aaron Trehub and Katherine Skinner

Getting to the Bottom Line: 20 Cost Questions for Digital Preservation is a cost-gathering resource created by the Outreach Committee of the MetaArchive Cooperative in Spring 2015. Launched during an Association of Southeastern Research Libraries (ASERL) webinar on March 11, 2015, this resource has been shared broadly with libraries, archives, and other institutions that have an interest in procuring digital preservation services. The easy-to-use resource is designed to equip institutions with questions that they can use to identify the full range of costs that might be associated with any particular digital preservation service--proprietary, community-sourced, or otherwise. For a variety of reasons, services of all types do not always make their costs as transparent as institutions might prefer. Using the Getting to the Bottom Line question-set will help ensure that institutions do not leave any stones unturned when evaluating their options and that they gather the information that they need to make informed choices that lead to sustainable solutions. Institutions are encouraged to make free use of the questions, adapt them as needed, and provide feedback on their usefulness. Going forward, the resource will serve as a foundation for building additional and more sophisticated cost transparency resources targeted toward the digital preservation community.

Innovative Session

Preservation Storage Community Discussion
Facilitator: Eld Zierau

Preservation storage is a core component of a sustainable digital preservation program and many organizations are wading through available and emerging options, both locally and beyond. This facilitated community discussion will open with some examples organizational approaches, strategies, and possible services then pose a series of questions to help participants identify and weigh options in relation to requirements, available resources, compliance, and feasibility. Note takers will capture highlights and outcomes from the discussion to share following the session. Please do bring along (and/or tweet) your examples and questions!

12pm-1pm **LUNCH**

1pm-2:30pm **SESSIONS**

Panel

Long-Term Preservation Strategies & Architecture: Views from Implementers. Mary Molinaro, Katherine Skinner, Sibyl Schaefer, Dave Pcolar, and Sam Meister

Join us for a panel presentation on the dark side of preservation. This panel will address the current state of long-term digital preservation and where we've come in the last decade. The presenters will review the tools and techniques for their projects and how they work. The panel will engage in an open discussion on the issues around long-term digital preservation, including: costs, technology, hurdles (technical and political) and planning for the future. The panel will also address how long-term digital preservation transcends disciplinary boundaries of librarianship and computer science and what values are implicit in the work and activities.

Innovative Session

Policy and Practice Documentation Clinic
Facilitator: Maureen Pennock

To demonstrate good practice, digital preservation programs need to develop, accumulate, preserve and make available as appropriate relevant and requisite policies and related documentation – evidence that repositories are addressing the set of emerging and evolving standards and requirements. This informal session will open with an overview of some organizational examples then break into small group to review examples and address questions from participants. Come ready with (and/or tweet) your examples and questions!

1pm-2:30pm **Panel**

Preservation of Research Data for Reuse. Ixchel Faniel, Elizabeth Hull, Vessela Ensberg, Seth Shaw and Reagan Moore

This panel aims to link research and practice around the preservation necessary for meaningful reuse of research data over the long term. Panelists will discuss preserving the contexts around the meaning of data that enable assessments of data quality necessary for reuse, preserving the bits of data that enable long-term access across the continuum and rendering, and shaping research data services to address the two in a more effective, integrated manner.

2:30pm-3pm **BREAK**

3pm-5:30pm **POSTER AND DEMO SESSIONS**

Demos/Posters

Poster/Demo Lighting Talks

1. **Mind the Gap. Bridging Digital Libraries & Archives**
   Mark Leggott and Erin Tripp
2. **Managing and Preserving Research Data in Ex Libris Rosetta**
   Adi Alter and Ido Peled
3. **The Oracle Cloud Storage Archive for Long-term Storage and Preservation**
   Pyounguk Cho and Art Pasquinelli

Infrastructure and Community Standard of Digital Preservation

4. **Strategies for Audit-based Repository Certification: Guidelines, Resources, and Tools to Prepare, Organize, and Evaluate Criteria Evidence**
   Jessica Tieman
5. **In the Thicket of It with the NDSA Standards and Practices Working Group: Cultivating Grass Roots Approaches to Real-World Digital Preservation Issues**
   Winston Atkins, Erin Engle, Andrea Goethals, Karl Jackson, Kate Murray, Carol Kussmann, Michelle Paolillo and Mariella Soprano
6. **Alternatives for Long-Term Storage Of Digital Information**
   Chris Erickson and Barry Lunt
7. **Invitation to Join the OAIS Community Platform**
   Barbara Sierman, William Kilbride, Hervé L'Hours and Paul Wheatley
8. **Open Preservation Foundation Community Survey 2015**
   Ed Fay, Becky McGuinness, Carl Wilson and Nick Krabbenhoeft
9. **Addressing Major Digital Archiving Challenges**
   Janet Delve, David Anderson and Andrew Wilson

Innovation in Workflow and Practice

11. **Automatic Identification and Preservation of National Parts of the Internet Outside a Country's Top Level Domain**
    Eld Zierau
12. **An Institutional Digital Repository Backbone**
    Adi Alter and Ido Peled
13. **ArchivesSpace-Archivematica-DSpace Workflow Integration**
    Michael Shallcross and Max Eckard
14. **Dash Curation Service Infrastructure Enhancement: An Informed Extension & Redesign**
    Nancy Hoebelheinrich and Stephen Abrams
15. **Minimal Effort Ingest**
    Bolette Ammitzbøll Jurik, Asger Askov Blekinge and Kåre Fiedler Christiansen

Emulation, Migration, and Long-term Accessibility

Teaching, Education, and Research Data Management

34. **Targeting Audiences among the Masses: A Data Curation MOOC for Researchers and Information Professionals**
    Helen Tibbo, Thu-Mai Christian and Rachel Goatley
35. **Preserving Electronic Syllabi at California State University Long Beach**
    Chloé Pascual
36. **What We Teach: An Assessment of Graduate-Level Digital Curation Syllabi**
    Carolyn Hank, Noah Lasley, Xiaohua Zhu, Kylan Shireman and Charlene N. Kirkpatrick
37. **Congregating Socio-Economic Data Sets for Scholastic Research: A Case Study in IIMB Library**

5:30pm-8pm **National Digital Stewardship Alliance (NDSA) Awards Reception**

Reception Sponsored by Digital Preservation Network
Best Poster Award Sponsored by School of Information and Library Science,
University of North Carolina at Chapel Hill
Best Paper Award Sponsored by Ex Libris

Performance by The Carolina Heartland Cloggers

Founded in 1984, the Carolina Heartland Cloggers, are an adult traditional clogging team that exhibits a variety of styles which exemplify the rich heritage and art of Southern Appalachian clogging in North Carolina. Their dance routines include freestyles, precision, smooth, hoedown, show and line.

Thursday, November 5, 2015

8am-5pm **Registration**

9am-10am **KEYNOTE**

Mass Digitization of Cultural Heritage: Can Copyright Obstacles Be Overcome?
Pam Samuelson, Richard M. Sherman Distinguished Professor of Law; Professor of School Information; Co-Director, Berkeley Center for Law & Technology

Chair: Anne Gilliland, Scholarly Communications Officer and Associate Law Librarian, University of North Carolina at Chapel Hill

Preserving cultural heritage is an important obligation that our society owes to future generations. Digital technologies have opened up new opportunities for engaging in preservation activities. Copyright is sometimes a significant impediment to digital preservation, although to be sure, it is far from the only challenge digital preservationists face. This talk will focus attention on the role that fair use may play in surmounting the copyright challenges, in light of the very recent appellate court decision in Authors Guild v. Google Inc. The Authors Guild v. HathiTrust appellate court decision from the previous year has affirmed the fairness of digitizing for purposes of creating a full-text searchable database, preserving in-copyright materials, and enhancing access to the contents of books for print-disabled persons. The Google decision makes it clear that serving up snippets that do not show enough of the expression in copyrighted materials to supplant market demand is fair use. Although the Authors Guild has announced that it will ask the Supreme Court to review the Google decision, this talk will explain why I think that appeal will not be successful. The greater challenge, however, is how to increase public access to the contents of the cultural artifacts of the 20th century beyond snippets. This talk will consider how much work fair use can do to achieve this objective and will discuss the Copyright Office's proposal for an extended collective license solution to the problem of attaining more access to the contents of in-copyright materials.

10am-10:30am **Break**

10:30am-12pm

Digital Preservation Showcase

For the showcase, representatives of four preservation software environments - Archivematica, Islandora, Preservica, and Rosetta - will explain how their software addresses specific points in three main functional areas: ingest, preservation/ storage, and access
Presenters:
- Archivematica: Evelyn McLellan (President, Artefactual)
- Islandora: Mark Leggott (Founder and Treasurer, Islandora)
- Preservica: Michael Hope (Senior Technology Marketing Manager)
- Rosetta: Christa Jameson (Meetings and Events Manager, Ex Libris)

Session 1 (10:30am-12:00pm) – Ingest (Moderator, Janet Delve, Creative Technologies)

Get a Room

Do you have an idea for a session for iPres 2016? Do you want to brainstorm with colleagues about a possible collaborative project? Do you want to continue a discussion of a topic raised in a session during the week? Sign up to get a room – sign-up sheets will be available from Tuesday morning until the end of lunch on Wednesday. Be sure to vote for a session if you're interested in participating. Room assignments will be announced at the National Digital Stewardship Alliance (NDSA) Awards Reception on Wednesday evening.

12pm-1pm

LUNCH

1pm-3:30pm

Digital Preservation Showcase

1:00pm-2:00pm - Session Two
Preservation and Storage (Moderator, Carl Wilson, Open Preservation Foundation)
2:00pm-3:00pm - Session Three
Access (Moderator, Carl Wilson, Open Preservation Foundation)
3:00pm-3:30pm - Wrap-Up
Open Discussion (Moderator, Kam Woods, University of North Carolina)

Get a Room

Do you have an idea for a session for iPres 2016? Do you want to brainstorm with colleagues about a possible collaborative project? Do you want to continue a discussion of a topic raised in a session during the week? Sign up to get a room – sign-up sheets will be available from Tuesday morning until the end of lunch on Wednesday. Be sure to vote for a session if you're interested in participating. Room assignments will be announced at the National Digital Stewardship Alliance (NDSA) Awards Reception on Wednesday evening.

3:30pm-4pm

Break

4pm-5pm

Watch This Space: What's Happening in the Digital Preservation Community that Might Become Next Year's Highlights? Facilitator: Leo Konstantelos

"What would you recommend to others in the digital preservation community to watch over the next year?" Paralleling the Spotlight portion of the Opening Session, the facilitator and presenters will suggest their own examples of project, initiatives, and developments to watch over the next year. The session will reference examples from presenters during the week and build on examples provided contributed by attendees and in tweets from across the digital preservation community.

Friday, November 6, 2015
Workshops & Tutorials

Note: All Friday workshops and tutorials will be held on the campus of the University of North Carolina at Chapel Hill, not the Friday Center.
Lunch for all Friday workshops and tutorials will be in Lenoir Hall, Mainstreet. You will receive a voucher to purchase lunch if you have registered for one of these workshops.

8am-5pm **REGISTRATION**

9am-12pm **PREMIS Implementation Fair**

Evelyn McLellan (Artefactual Systems), Karin Bredenberg (National Archives of Sweden) and Rebecca Guenther (Consultant and Library of Congress)

This workshop provides PREMIS implementers with an overview of the changes in the PREMIS Data Dictionary for Preservation Metadata, version 3.0. As an international standard for metadata to support the digital preservation process, PREMIS has been implemented world-wide and is incorporated in many commercial and open-source digital preservation tools and systems. With the release of version 3.0 in June 2015, implementers have enhanced ability to describe their digital assets, including a new way of describing complex software and hardware environments that are so important to their preservation and future use. There will also be a report on the integration of preservation systems and tools that provide different functions in management and preservation. Implementers are encouraged to report on their experiences using PREMIS, particularly issues encountered, and there will be ample time for discussion.

8am-5pm **Using Open-Source Tools to Fulfill Digital Preservation Requirements**

Courtney Mumma (Internet Archive), Bradley Westbrook (Lyrasis), Michael Shallcross (University of Michigan), Sam Meister (Educopia), Christine Di Bella (Lyrasis), Max Eckard, (University of Michigan) and Christopher (Cal) Lee (University of North Carolina)

This workshop offers a space to talk about open-source software for digital preservation, and the particular challenges of developing systems and integrating them into local environments and workflows. Topics will include current efforts and grant-funded initiatives to integrate different open source archival software tools; the development of workflows involving multiple open source tools for digital preservation, forensics, discovery and access; and the identification of gaps which may need filled by these or other tools.

Curating Research Assets and Data Using Lifecycle Education

Helen Tibbo (UNC – Chapel Hill) and Thu-Mai Christian (UNC – Chapel Hill)

As major funding agencies, publishers, and research institutions continue to issue data sharing, management, and archiving policies in increasing numbers, libraries are being called upon to support researchers in their efforts to comply with these policies. To be responsive to researchers' data needs and to increase the likelihood of effective and efficient data preservation, many data librarians and archivists are seeking the knowledge, skills, and competencies necessary to confront the growing— and increasingly complex—data management and preservation needs of their institutions. With lecture, discussion, and hands-on exercises, this tutorial will explore the obligations of researchers to manage their data, identify the attributes of data that add to the complexity of data curation tasks, and introduce a range of tools and resources available to help librarians effectively implement data curation, and particularly, preservation services.

Benchmarking Forum

Kresimir Duretec (Technical University of Vienna), Artur Kulmukhametov (Technical University of Vienna), Christoph Becker (University of Toronto , Technical University of Vienna) and Andreas Rauber (Technical University of Vienna)

The quality of digital preservation tools is of great importance to the preservation community. However, quality assessment is often done in an isolated way with a lack of systematic and community driven initiatives. Benchmarking is a method of comparing entities to a well-defined standard (benchmark) that has shown itself as a valuable empirical method for evaluating software tools. The successfulness of benchmarking is dependent on the readiness of the community to accept and drive the whole process. This workshop is focused on discussing software benchmarking practices in digital preservation and how these can contribute to improving digital preservation tools.

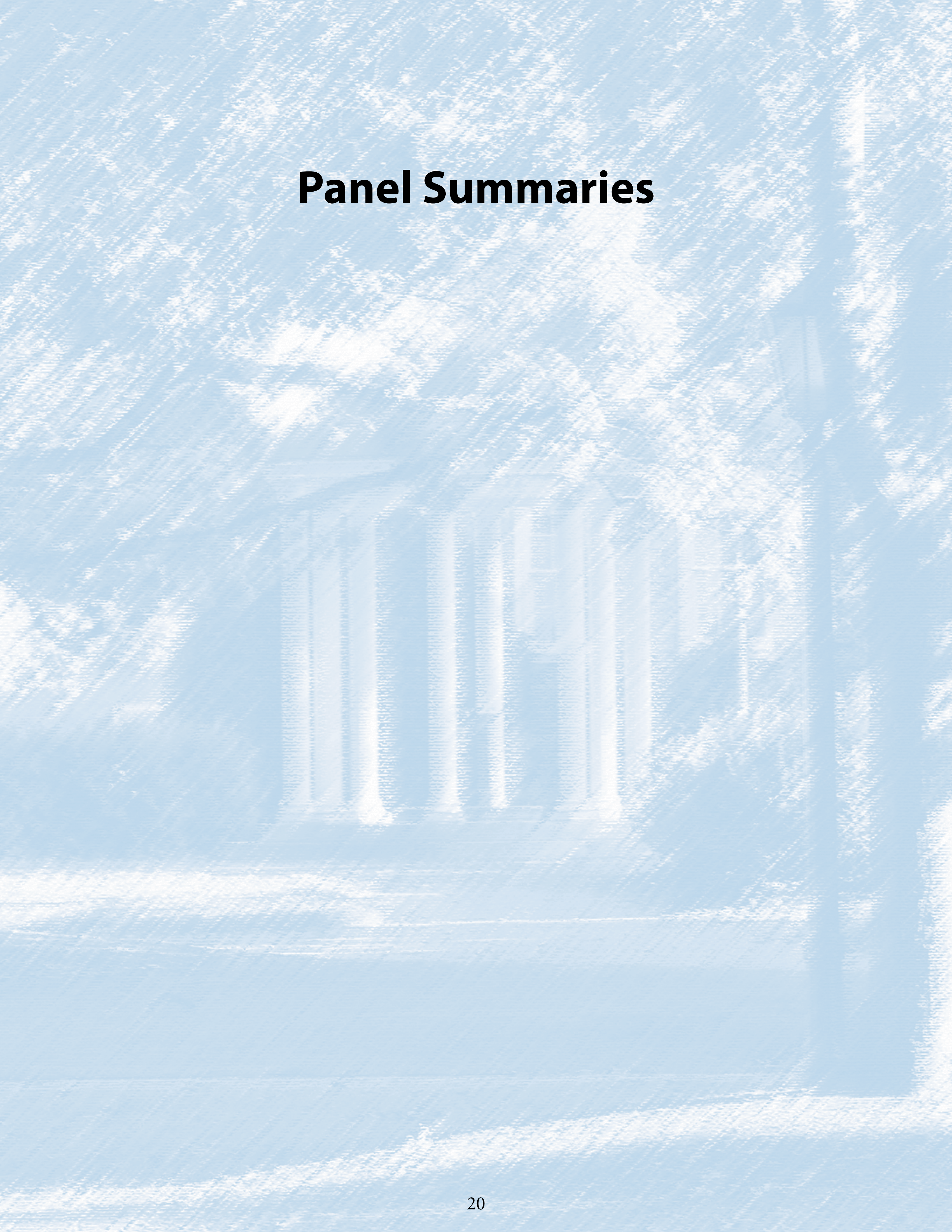
1pm-5pm

Data Mining Web Archives

Jefferson Bailey (Internet Archive)

This workshop will explore new methods of research use of web archives by giving attendees exposure to, and training in, the tools, methods, and types of analysis possible in working with datasets extracted from the entirety of curated web archive collections. Giving researchers datasets of specific extracted metadata elements, link graph data, named entities, and other post-processed data can help facilitate new uses and new types of visualization, inquiry, and analysis.

Panel Summaries



Advancing the Evidence Base of Digital Preservation

Micah Altman
MIT Libraries
Massachusetts Institute of Technology
E25-131 (45 Carlton Street)
Cambridge, MA
+1-585-466-4224
escience@MIT.EDU

Jonathan Crabtree
Odum Institute
UNC - Chapel Hill
229B Davis Library CB#3355
Chapel Hill, NC 27599-3355
+1-919-962-0517
jonathan_crabtree@unc.edu

Helen R. Tibbo
School of Info. & Lib. Science
UNC - Chapel Hill
201 Manning Hall CB#3360
Chapel Hill, NC 27599-3360
+1-919-962-8063
Tibbo@ils.unc.edu

ABSTRACT

Research is critical to the advancement of both a basic understanding and the effective practice of digital preservation. R esearch must, however, be intimately linked to practice in order to improve outcomes. This panel will discuss methodology, metrics, tools, and exemplars that can effectively build the evidence base for digital preservation. Panelists will present on a simulation framework for evaluating preservation risks, formal/machine actionable preservation strategies and implementations; and evaluation of preservation performance.

General Terms

Innovative practice; Training and education.

Keywords

Methodology, Research, Metrics, Experiments.

1. INTRODUCTION

Research is critical to the advancement of both a basic understanding and the effective practice of digital preservation. But research must be intimately linked to practice in order to improve outcomes. This panel will discuss methodology, metrics, tools, and exemplars that can effectively build the evidence base for digital preservation. Panelists will address the following questions:

- What exemplars have been most successful in systematically contributing to the overall cumulative evidence base for digital preservation practice?
- How can replicable, scalable research and assessment methods -- including trend analysis, simulation, and designed experiments be best integrated into preservation practices?
- What approaches are most successful in integrating research and practice (research-based-practice or practice-based-research)?
- What are the strengths and weakness of current metrics and measurements for digital preservation practice?

Micah Altman will discuss a simulation framework for evaluating preservation risks. Jonathan Crabtree will present on formal/machine actionable preservation. Finally, Helen Tibbo will discuss the development of ISO 16363, its subsequent use to ensure high quality repositories, and the potential for research using ISO findings.

2. ACKNOWLEDGMENTS

Our thanks to the National Digital Stewardship Alliance for their support in developing this panel.

Preservation of Research Data for Reuse

Ixchel M. Faniel
OCLC
6565 Kilgour Place
Dublin, OH 43017-3395
fanieli@oclc.org

Seth Shaw
Clayton State University
2000 Clayton State Blvd
Morrow, GA
sethshaw@clayton.edu

Elizabeth Hull
Dryad Digital Repository
PO Box 585
Durham, NC 27701
ehull@datadryad.org

Vessela Ensberg
UCLA Library 12-077
Center for Health Sciences
Los Angeles, CA 90095
vensberg@library.ucla.edu

Reagan Moore
UNC-Chapel Hill
108 Homewood Drive
Chapel Hill, NC 27514
rwmoore@email.unc.edu

ABSTRACT

This panel aims to link research and practice around the preservation necessary for meaningful reuse of research data over the long term. Panelists will discuss preserving the contexts around the meaning of data that enable assessments of data quality necessary for reuse, preserving the bits of data that enable long term access across the continuum and rendering, and shaping research data services to address the two in a more effective, integrated manner.

General Terms

Institutional opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows

Keywords

data reuse, preservation, research data services, digital curation

1. INTRODUCTION

Disciplinary researchers and technologists view the problem of reusing research data from different perspectives – preservation of the research context for meaning and preservation of the technological context for future 'performance' or rendering. In fact, there is little overlap in the literature examining these two perspectives, e.g., data reuse and data curation. The data reuse literature primarily focuses on the preservation of meaning that facilitates researchers' assessments of data quality that, in turn, enable reuse. Taking a user-centric approach, data reuse studies tend to identify the contextual information (i.e. significant properties) necessary to help people assess whether data are relevant, credible, interpretable, and trustworthy [2, 3, 10, 12]. In contrast, the data curation literature tends to take a data-centric approach to identify the significant properties that support long term reliable access to digital resources across the continuum in order to maintain data's functionality, appearance, and computing environment [1, 4, 6, 9, 11]. These perspectives are not in opposition but exist along a scale. Both are necessary and a balance between the two is an imperative. This is particularly evident in the work of data librarians and digital archivists who occupy the space between data producers and repositories in an effort to ensure efficient and effective reuse.

When shaping data services, data librarians often find themselves negotiating between disciplinary researchers and repository managers. This is where the gaps between the contextual information researchers generate and use in the course of their daily work and the contextual information necessary in a repository to enable discovery and effective management of data resources becomes apparent [7, 8, 12]. As a result, data librarians often find themselves in the position of bridging between communities: looking for ways to make the process of externalization more attractive and useful to data producers and to broaden technologists' thinking around what is really needed to manage and preserve data across the continuum.

At the same time, data librarians and digital archivists are considering ways to reduce the time spent on preservation activities. Take format migration as an example. It is time consuming for all but the most well-defined formats, which researchers do not typically use or prefer. Developing and relying on international standards to enable automated format migration for a large variety of files would reduce some of the burden. However, it requires data professionals to work beyond the confines of their institutions and partner with external entities to speed up standards development.

This panel aims to link research and practice around the preservation of research data through various perspectives - researcher, librarian, repository staff, archivist, information scientist, instructor. The panel will focus on the different types of contextual information required for meaningful reuse over the long term, the technological context to ensure digital 'performance', and the intermediary people, practices, and services required to ensure that it is obtained. Each panelist will take 5-10 minutes to introduce their perspectives on or approaches to the preservation of research data for reuse. Their introductions will be followed by a moderated discussion with the audience.

2. PANEL PARTICIPANTS

Moderator: Arcot Rajasekar, Ph.D., is a Professor in the School of Information and Library Sciences at the University of North Carolina at Chapel Hill, a Chief Scientist at the Renaissance Computing Institute (RENCI), and a Co-Director of Data Intensive Cyber Environments (DICE) Center at UNC. A leading proponent of policy-oriented, large scale data management, Rajaseker has several research projects and over 150 publications in the areas of data grids, digital libraries, persistent archives, logic programming and artificial intelligence.

Panelist: Ixchel M. Faniel, Ph.D., is a Research Scientist at OCLC. Faniel's current work examines data reuse practices in several disciplinary communities and academic librarians' experiences developing and delivering research data services. Faniel will discuss findings from a comparative study of data reuse practices in three disciplinary communities and highlight the significant properties of data across the disciplines that facilitate the preservation of meaning necessary for data reuse (http://dipir.org).

Panelist: Seth Shaw, MSI, is an Assistant Professor of Archival Studies at Clayton State University. His focus is on teaching archival theory and practice with an emphasis in the implications of modern technology. Shaw will describe the placement of preservation practices on a scale of context, representation, and meaning from the technical to the conceptual level with an emphasis on the adaptive and secondary performances required for research data reuse [5]. He will also describe the pedagogical approach used while training digital archivists to convey a holistic understanding of digital content as layered representations with adaptable performances.

Panelist: Elizabeth Hull, MA, is Operations Manager for Dryad, an independent, nonprofit digital repository for data underlying the scientific and medical literature. As part of her role, Hull facilitates data curation and oversees the repository helpdesk. Hull will address Dryad's challenges in balancing preservation and reuse while trying to keep the burden of data archiving as low as possible for researchers. She will share some of Dryad's experiences in working to encourage good documentation and retain usefulness of Dryad data packages into the future.

Panelist: Vessela Ensberg, Ph.D., is a Data Curation Analyst at the UCLA Louise M. Darling Biomedical Library and at the UCLA Data Archive. Working at both departments she has the opportunity to work with data throughout the lifecycle from planning to preservation. Ensberg will discuss her work on a project to enrich the PRONOM file registry with information on files that researchers use. Her goal is to help speed up the automation of file format migration.

Panelist: Reagan Moore, Ph.D., is a Professor in the School of Information and Library Science at University of North Carolina at Chapel Hill. His research interests are on policy-based data management systems. Moore leads the Data Intensive Cyber Environments Center at UNC, which develops the integrated Rule Oriented Data System. The software is used to manage archive, digital libraries, and research collaboration environments. Moore will discuss preservation policies for research data, and the workflows used to generate the data. For reproducible research, a future researcher should be able to re-execute the analysis and generate the same result [8].

3. ACKNOWLEDGMENTS

The DIPIR Project (http://dipir.org) was made possible by a National Leadership Grant from the Institute for Museum and Library Services (IMLS), LG-06-10-0140-10, "Dissemination Information Packages for Information Reuse."

4. REFERENCES

[1] Coyne, M., Duce, D., Hopgood, B., Mallen, G., and Stapleton, M. 2007. The significant properties of vector images. Available at http://www.jisc.ac.uk/media/documents/programmes/preservation/vector_images.pdf.

[2] Faniel, I., Kansa, E., Whitcher Kansa, S., Barrera-Gomez, J., and Yakel, E. 2013. The challenges of digging data: A study of context in archaeological data reuse. In *Proceedings of the Joint Conference on Digital Libraries* (Indianapolis, IN, July 2013), ACM, 295-304.

[3] Faniel, I. M., Kriesberg, A., and Yakel, E. 2012. Data reuse and sensemaking among novice social scientists. *Proceedings of the Association for Information Science and Technology (ASIS&T)*. 49, 1, 1-10.

[4] Hedstrom, M., Lee, C., Olson, J., and Lampe, C. 2006. "The old version flickers more:" Digital preservation from the user's perspective. *AM Archivist*. 69, 1, 159–187.

[5] Heslop, H., Davis, S., and Wilson, A. 2002. *An Approach to the Preservation of Digital Records*. Canberra: National Archives of Australia. Available at http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm16-47161.pdf.

[6] Matthews, B., McIlwrath, B., Giaretta, D., and Conway, E. 2008. *The Significant Properties of Software: A Study*. Rutherford Appleton Laboratory: Joint Information Systems Committee. Available at http://www.jisc.ac.uk/whatwedo/programmes/preservation/2008sigprops.

[7] Mayernik, M. S. 2010. Metadata tensions: A case study of library principles vs. everyday scientific data practices. *Proceedings of the American Society for Information Science and Technology*, *47*, 1, 1–2.

[8] Moore, R. A., and Rajasekar, H. Xu. 2015. *DataNet Federation Consortium Policy Toolkits*. iPRES Conference, November 2015.

[9] Morrissey, S. 2010. The economy of free and open source software in the preservation of digital artifacts. *LIBR HI TECH*. 28, 2, 211-223.

[10] Rolland, B., and Lee, C. P. 2013. Beyond trust and reliability: Reusing data in collaborative cancer epidemiology research. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, TX, February 2013). ACM, 435-444.

[11] Rosenthal, D. S. H. 2010. Bit preservation: A solved problem? *International Journal of Digital Curation*. 5, 1, 134-148. DOI=10.2218/ijdc.v5i1.148

[12] White, H. C. 2014. Descriptive metadata for scientific data repositories: A comparison of information scientist and scientist organizing behaviors. *Journal of Library Metadata*. 14, 1, 24–51.

[13] Zimmerman, A. S. 2008. New knowledge form old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology & Human Values*. 33, 5, 631-652.

Preserving Born-Digital News

Edward McCain
Digital Curator for Journalism
Donald W. Reynold Institute
218 RJI, Missouri School of Journalism
Columbia, MO 65211
mccaine@rjionline.org

Hannah Sommers
Associate University Librarian
The George Washington University
2130 H St. NW, Ste 606E
Washington, DC 20052

Martin Klein
Programmer/Analyst
University of California Los Angeles
Research Library
Los Angeles, CA
Martinklein0815@gmail.com

Christie Moffatt
Digital Manuscripts Program
History of Medicine Division
National Library of Medicine
Bethesda, MD
moffattc@mail.nlm.nih.gov

Abigail Potter
National Digital Information Infrastructure and Preservation Program
Library of Congress
Washington, DC
abpo@loc.gov

Stéphane Reecht
Digital Curator
Product Owner of SPAR
Preservation and Curation Department
Bibliothèque Nationale de France
T1 N7 10. Quai François Mauriac, 75706
Paris Cedex 13
stephane.reecht@bnf.fr

ABSTRACT

The news industry has quickly adopted networked digital technologies to create and distribute their content across all media types and in an ever-increasing number of formats. These technologies have also enabled individuals to capture and share information, news, and opinion on contemporary and community events. These changes contribute to a dynamic news ecosystem, upending traditional publishing models that media companies, libraries, archives and memory institutions have depended on to save the news. In this panel, the challenges and opportunities of preserving born-digital news content will be presented and discussed. A preliminary environmental scan of the state of digital news preservation will be shared. Perspectives and tactics from the "front-line" of news creation will be covered in addition to establishing special collections to capture and preserve web sites that cover news events. Efforts to establish relationships with the creators of content management systems (CMS) that drive the back end of modern media publishing networks will also be presented, as will tools that have been developed to capture social media and other content from the web that contributes to the present day news ecosystem.

ACKNOWLEDGMENTS

Outcomes and results from the two Dodging the Memory Hole events contributed to this panel. Panelists are members of and collaborate via the National Digital Stewardship Alliance Content Working Group.

Long Term Preservation Strategies & Architecture: Views from Implementers

Mary Molinaro, Dave Pcolar
Digital Preservation Network
mary, dave {@dpn.org}

Katherine Skinner, Sam Meister
Educopia Institute
katherine, sam {@educopia.org}

Sybil Schaefer
UCSD Libraries/Chronopolis
sschaefer@ucsd.edu

ABSTRACT

Join us for a panel presentation on the dark side of preservation. This panel will address the current state of long-term digital preservation and where we've come in the last decade. The presenters will review the tools and techniques for their projects and how they work. The panel will engage in an open discussion on the issues around long-term digital preservation, including: costs, technology, hurdles (technical and political) and planning for the future. The panel will also address how long-term digital preservation transcends disciplinary boundaries of librarianship and computer science and what values are implicit in the work and activities.

Good, Better, Best? Examining the Range and Rationales of Institutional Data Curation Practices

Robin Rice
Data Librarian, EDINA and Data Library,
University of Edinburgh
160 Causewayside, Edinburgh EH9 1PR
UNITED KINGDOM
+44 131 651 1317
r.rice@ed.ac.uk

Limor Peer
Associate Director for Research, Institution for Social and Policy Studies, Yale University
77 Prospect Street, New Haven, CT 06520-8209
UNITED STATES
+1 203 432 0054
limor.peer@yale.edu

Wendy White
Head of Scholarly Communication
University of Southampton
University Road, Southampton SO17 1BJ
UNITED KINGDOM
+44 (0)23 8059 6873
whw@soton.ac.uk

Florio Arguillas
Research Associate, Cornell Institute for Social and Economic Research (CISER), Cornell University
391 Pine Tree Rd., Ithaca, NY 14850
UNITED STATES
+1 607 255 7838
foa2@cornell.edu

ABSTRACT

Many academic institutions are grappling with managing local research data assets. Resources and approaches vary. This panel will explore curation procedures at institutional data repositories.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges; Preservation strategies and workflows.

Keywords

Data curation, Data management, Institutional repositories

1. THEME

Many academic institutions are grappling with managing local research data assets for the first time, due to recent demand from users and research funders; others have been providing forms of data access and support for years, allowing them to gradually build up knowledge and applied expertise. Similarly, the resources available for data archiving and support vary enormously between institutions, based on perceived importance of research data management (RDM) by senior managers, degree of commitment to long-term preservation and reuse of the data, the level of funding brought in by research activity, the extent of support provided in-house by libraries and IT centres, relative size of the institution, and number of disciplines in scope.

A wide range of technological solutions are available to data archivists, who come from diverse backgrounds, profess varied academic and professional values in relation to 'open knowledge' and data sharing, and may be located in various units within the university. Finally, knowledge of digital curation and preservation standards and techniques is not uniform amongst practitioners.

This panel, like the IPRES 2015 delegation, reflects a portion of that range of variation in terms of approaches to data archiving, curation, digital preservation, and support for data re-use. The panel will offer an unabashed look at today's approaches to institutional data curation that weighs 'best' practice against real world concerns for conservation of resources and meeting expectations of local and other stake-holders.

2. PROGRAM

Time	Subject	Presenter
15 minutes	Introduction Good, better, best? Examining the range and rationales of institutional data curation practices	Robin Rice
45 minutes	Panel discussion Each panel member will describe their own data curation and archiving approaches based on comparative topics such as, types of data received and issued, repository governance, mission and policies, technological choices, levels of engagement with depositors and end-users, degree of elaboration for workflows for data management and preservation planning, quality assurance of content (from minimal to maximal), and appraisal protocols.	Panelists Robin Rice Limor Peer Wendy White Florio Arguillas
30 minutes	Audience Q&A with the panel	Moderated by Limor Peer

3. PRESENTERS AND PANELISTS

Robin Rice, Data Librarian, manages the Data Library services at the University of Edinburgh, including Edinburgh DataShare, an institutional data repository, and Research Data MANTRA, an open online training course in data management. She is active in the University's Research Data Management programme. She received a BA at Clark University and a Masters in Library and Information Studies from the University of Wisconsin-Madison.

Limor Peer, PhD, is Associate Director for Research, Institution for Social and Policy Studies, Yale University. She oversees research infrastructure and process at ISPS and manages a specialized research data repository (the ISPS Data Archive) and is currently involved in campus-wide efforts relating to research data sharing and preservation. Limor received a Ph.D. and M.A. in Communication Studies from Northwestern University and a B.A. in Political Science from Tel-Aviv University.

Wendy White is Head of Scholarly Communication at the University of Southampton where her work includes leading the co-ordination of cross-service support for the curation and discovery of all types of research output. She has been involved with a range of projects relating to innovations supporting open access, repositories and research data, including as Principle Investigator for the Jisc funded institutional DataPool project (2011-13) and Co-Investigator on IDMB (2009-11) which developed a 10 year strategic roadmap for research data management at Southampton.

Florio Arguillas, Jr. is Research Associate at the Cornell Institute for Social and Economic Research where his work involves, among others, managing the CISER Helpdesk, which provides a whole range of free statistical software consulting services to the Cornell community; managing CISER's Data Curation and Replication Service; and co-managing the data curation, management, and archiving training offered by CISER. Florio received an AB in Economics from Ateneo de Davao University, M.A. in Demography from the University in the Philippines, and an M.S. and Ph.D. in Development Sociology from Cornell University.

Engaging Content Creators to Improve the Capture and Preservation of Born-Digital Content

Imogen Smith
Project Manager
Dance Heritage Coalition
Washington, DC
ismith@danceheritage.org

David Walls
Preservation Librarian
Government Printing Office
Washington, DC
dwalls@gpo.gov

Martin Halbert
Dean of Libraries
University of North Texas
Denton, TX
Martin.Halbert@unt.edu

Abigail Potter
National Digital Information Infrastructure and Preservation Program
Library of Congress
Washington, DC
abpo@loc.edu

ABSTRACT

The 2015 *National Agenda for Digital Stewardship* calls for stewardship organizations to engage content creators to improve the capture and preservation of born-digital content. This panel will provide an overview of three different efforts to reach out to different communities who create content that stewardship organizations want to acquire and preserve. The Dance Heritage Coalition will present outcomes from their Knight funded project to work directly with dance companies and critics to capture born-digital content documenting dance and dance performance. Efforts of the U.S. Federal Web Archiving Working Group to interact with federal webmasters to improve the capture and preservability of federal information and government web sites will also be covered. Finally, building relationships with community news media and developing donor agreements for the transfer of assets will be discussed. Other born-digital content areas could also be covered.

ACKNOWLEDGMENTS

Panelists are members of and collaborate via the National Digital Stewardship Alliance Content Working Group.

Papers

A Foundational Framework for Digital Curation: The Sept Domain Model

Stephen Abrams
California Digital Library
University of California
Oakland, CA 94612, US
+1 510-987-0370
Stephen.Abrams@ucop.edu

ABSTRACT

Digital curation is a complex of actors, policies, practices, and technologies enabling successful consumer engagement with authentic content of interest across space and time. While digital curation is a rapidly maturing field, it still lacks a convincing unified theoretical foundation. A recent internal evaluation by the University of California Curation Center (UC3) of its programmatic activities led quickly to seemingly simple, yet deceptively difficult-to-answer questions. Too many fundamental terms of curation practice remain overloaded and under-formalized, perhaps none more so than "digital object." To address these concerns, UC3 is developing a new model for conceptualizing the curation domain. While drawing freely from many significant prior efforts, the UC3 Sept model also assumes that digital curation is an inherently semiotic activity. Consequently, the model considers curated content with respect to six characteristic dimensions: semantics, syntactics, empirics, pragmatics, diplomatics, and dynamics, which refer respectively to content's underlying abstract meaning or emotional affect, symbolic encoding structures, physical representations, realizing behaviors, evidential authenticity and reliability, and evolution through time. Correspondingly, the model defines an object typology of increasing consumer utility and value: blobs, artifacts, exemplars, products, assets, records, and heirlooms, which are respectively existential, intentional, purposeful, interpretable, useful, trustworthy, and resilient digital objects. Content engagement is modeled in terms of creator, owner, curator, and consumer roles acting within a continuum of concerns for catalyzing, organizing, and pluralizing curated content. Content policy and strategy are modeled in terms of seven high-level imperatives: predilect, collect, protect, introspect, project, connect, and reflect. A consistent, comprehensive, and conceptually parsimonious domain model is important for planning, performing, and evaluating programmatic activities in a rigorous and systematic rather than ad hoc or idiosyncratic manner. The UC3 Sept model can be used to make precise yet concise statements regarding curation intentions, activities, and results.

General Terms

Frameworks for digital preservation.

Keywords

Digital curation, digital preservation, domain model, semiotics, continuum, policy, strategy.

1. INTRODUCTION

Digital curation is a complex of actors, policies, practices, and technologies enabling successful consumer engagement with authentic content of interest across space and time.

A given unit of content is of *interest* if it can be readily distinguished from the larger universe of potential alternative content on the basis of consumer criteria, and *authentic* if it is what it purports to be. A consumer's engagement is *successful* if the content can be feasibly exploited for use and that use is beneficial for some desired purpose, ideally at a time and place and in a manner of the consumer's choosing. Feasibility of use depends upon intellectual and technical considerations regarding production and management, for example, selection, acquisition, arrangement, integrity, permission, visibility, etc., while the benefit of use is conditioned by individualistic purpose. It is possible that this purpose may be fulfilled only at some considerable spatio-temporal distance from the point of the content's creation; regardless, the consumer's purpose, and derived benefit, is not necessarily constrained to conform to the original intention of the content's creator, owner, or steward. Rather, every engagement is uniquely situated with respect to the context of the content's production, its curatorial framing, and its consumer's collateral experience, expertise, and expectation. Although this context is ultimately subjective, it may nevertheless be commonly held by other consumers participating in the same domains of discourse.

The curation attributes of enablement, success, engagement, authenticity, and interest are a contemporary restatement of traditional content stewardship concerns as articulated, for example, by Ranganathan's "laws" of library science [29]. The first law, "Books are for use," shorn of its biblio-centricity, is fundamentally concerned with *utility*, that is, the use for purpose underlying any successful engagement with a message-bearing object. The second and third laws, "Every reader his book" and "Every book its reader," are fundamentally concerned with ensuring an effective *connection* between content and consumer. The question of whether the "book" is what it purports to be is one of *authenticity*, a traditional concern of archival diplomatics that is especially important in the digital realm given content's ease of mutability. Mutability of a different sort is implicated in Ranganathan's fifth law, "The library is a growing organism," which is fundamentally concerned with *change*, corresponding to curation concerns with content's extension across space and time. The fourth law, "Save the time of the user," is fundamentally concerned with convenience, or more generally, *service*, and corresponds to the imperative of curating agents providing their customers with tools and services that effectively and efficiently meet their intellectual, behavioral, and technical expectations. Underlying all of these concerns is the notion that curation encompasses both preservation and use [42] [33], which are complementary rather than disparate activities: preservation ensuring use *over* time while use is dependent upon preservation *up until a point* in time.

Curation outcomes naturally lie along a spectrum of possible results largely dependent upon the degree to which appropriate human, organizational, and technical resources can be applied. Some of the factors pertinent to resource allocation decisions are intrinsic to the content itself, such as size, format, structure, and presence (or absence) of self-describing metadata; others are extrinsic, such as collection development policies, curatorial assessments of value, degree of uniqueness or ubiquity, ease of reacquisition or regeneration, availability of appropriate standards, best practices, and tools, staffing levels, and competing demands on finite organizational resources. Given the inevitability of resource constraints it is important that curating institutions make fully informed decisions to allocate (or withhold) resources and perform (or forgo) activities. This will enable institutions to plan and implement effective solutions that maximize curation utility, that is, provide the highest *overall* level of curation outcomes for the largest body of content with the least effort, while simultaneously expending *proportionate* effort towards any given unit or class of content based on its assessed value and institutional priority.

2. MODELING THE CURATION DOMAIN

Curation decisions should be made with respect to an underlying theory or conceptual domain model. A domain model is an abstraction of fundamental expressive and explanatory principles sophisticated enough to explicate past events and conditions and anticipate the consequences and efficacy of future decisions and actions; in other words, it should be both descriptive and predictive [30]. It is useful to build up such a model incrementally from first principles in order to ensure comprehensive scope, self-consistency, and conceptual parsimony. It is important, however, to keep in mind that all models are at best idealized representations of nominal domain concepts. The simplifying assumptions and abstractions inherent to any modeling effort may be at times incommensurate with pertinent real world detail and any actual curation entity or condition may not fully conform to model formalisms.

When the University of California Curation Center (UC3) first started a comprehensive internal review of its curation activities to evaluate their efficacy and set future priorities, it did so in the context of many descriptive and prescriptive frameworks familiar to the digital curation and preservation communities, for example, the ISO 14721 OAIS reference model, PREMIS, TRAC, etc. In working with these models, however, UC3 staff soon found themselves asking a number of seemingly simple, yet deceptively difficult-to-answer questions. What *exactly* is a "digital object"? (A bitstream? A file? A package? A dissemination?) What *specifically* is meant by "preservation" of an object? (A concern for the integrity of bits? Of context? Of performance? Of understanding?) None of the preexisting frameworks provided fully sufficient answers. In addition to definitional ambiguity, it was not immediately apparent how – or indeed whether – the conceptual models underlying these disparate efforts cohered into a unified and inclusive picture of the curation domain. A comprehensive reference model is important in ensuring that programmatic curation activities are planned, performed, and evaluated in a rigorous and systematic rather than ad hoc or idiosyncratic fashion. To address these concerns, UC3 has developed a new approach towards conceptualizing the curation domain that draws freely from past efforts, but also incorporates applicable concepts from other relevant fields such as information science, cognitive psychology, and semiotic theory. The UC3 Sept model affords a useful conceptual map, analytical framework, and descriptive vocabulary applicable to the full range of curation activities [41].

2.1 Curation semiotics

The ultimate goal of curation is to facilitate the effective "delivery" of content to human consumers across barriers imposed by space and time. (Even in cases of intermediating technical systems, ultimate agency always resides in a human actor [10].) In psycho-physiological terms, an act of content consumption occurs when:

1. An abstract unit of content is …
2. Realized by physical stimuli, which are …
3. Perceived by a sense modality, …
4. Interpreted in the specific subjective context of the consumer, and ultimately …
5. Experienced as cognitive meaning or psychological affect.

In making the final crucial transition from perception to cognition it is important to recognize that content consumption is an inherently semiotic act.

Semiotics is the study of signs and systems of signification, that is, things that *carry* messages and the ways in which those messages are represented and communicated [21] [25]. A sign is something that "stands in" for something else, in some manner, for someone [26]. In other words, it is a triadic relation between an external referent, its representation, and its effect on the consumer, which is a new mental state or reformulation of the referent stimulated by its representation. This cognitive or emotional state always arises in the subjective contextual ground of the consumer's collateral experience independent of the sign itself [5]. No unit of content is inherently significant; it gains significance for a given consuming agent only "in a context relevant to some purpose or goal" [12].

2.1.1 Roles

The consumer role is defined in the generic sense of an actor who derives some benefit from the direct use of, or indirect reference to, curated content. Direct exploitation may be consumptive (for example, passive reading, watching, listening to, etc.), generative (creating something new), or manipulative (adding to, modifying, or deleting from something extant). Indirect benefit, on the other hand, may be derived merely from the existence of content independent of direct use. The retention of certain legal materials, for example, confers tangible value to agents subject to relevant statutory or regulatory obligations or those with a legal interest in the materials' subject matter. The other fundamental curation roles are content creator, content owner, and content curator, corresponding to agents exercising creative, proprietorial, and delegated stewardship responsibilities, respectively. Any or all of these roles may be held by a given individual or corporate actor at various times and varying organizational and operational contexts.

2.1.2 Analytical concerns

For purposes of analysis, it is useful to consider digital content in terms of six characteristic dimensions: semantics, syntactics, empirics, pragmatics, diplomatics, and dynamics:

1. Semantics is concerned with the relationships between content and its underlying abstract meaning or affect;
2. Syntactics, with the relationships between content and its symbolic expressions;
3. Empirics, with the relationships between content and its physical representations [38];
4. Pragmatics, with the relationships between content and its consumers, that is, those concerning realizing

behaviors [24];

5. Diplomatics, with the relationships between content and the factual authenticity and reliability of its expression, representation, management, and transmission [32]; and
6. Dynamics, with the relationships between various states of content as it persists and evolves across space and time [9] [16].

(The term "diplomatics" is used here as a convenient generic label for a complex of concerns regarding trustworthiness rather than the more specific sense of use common to archival practice.)

These analytic dimensions correspond to longstanding stewardship concerns with content's abstract meaning, symbolic inner structure and outer form, physical carrier, behavioral experience, archival authenticity, and spatio-temporal persistence. They also give rise to the "Sept" model name, which was suggested by the approximate phonetic pronunciation of the SSEPDD dimensional acronym. "Sept" is also a genealogical term referring to a subgroup of an extended clan or family, appropriate for a model concerned with delineating nuanced distinctions within digital objects.

2.2 Object modeling

Digital objects are encapsulations of information for purposes of communication. Before devising Sept, UC3 reviewed a number of prior models for objects and the more general notion of communicable information, including the sender/receiver model (that is, Shannon and Weaver [36] as extended by Schram [35] and Berlo [6]); Buckland's information trichotomy [7]; Kahn and Wilensky [20], FRBR [18], the NAA performance model [17], OAIS [19], PREMIS [28], the Basic Representation Model (BRM) [43], and the Information Carrying Ontology (ICO) [14]. The component ontological subdivisions defined by these models can be approximately aligned against one another in a tabular fashion as shown in Table 1.Two pertinent facts emerge from this exercise: first, the number of rows in the table indicates the overall fineness of granularity with which these models have usefully decomposed the concept of an information object; and second, none of the prior efforts completely addresses the full gamut of ontological concerns at the finest decompositional level. The Sept model is intended to unambiguously defining of all ontological granules in a single coherent model, clarifying what an object is and what it is not.

Table 1. Information modeling crosswalk

Sender / receiver (1960)	**Buckland** (1991)	**Kahn-Wilensky** (1995)	**FRBR** (1998)	**NAA** (2002)	**OAIS** (2003)	**PREMIS** (2005)	**BRM** (2012)	**ICO** (2012)	**UC3 Sept** (2015)
source	info-as-knowledge	▨	work	essence	information object	intellectual entity	propositional content	intellectual entity	message
encoding	info-as-thing	data	expression	source	data object / digital object	bitstream / filestream	symbol structure	symbol structure	structure
			manifestation			file / representation			form
			item	▨	bits	▨	patterned matter/energy	information carrier	carrier
frame-of-reference		key-metadata	▨	▨	representation information	▨	auxiliary information	▨	annotation
channel	info-as-process	▨	▨	process		▨	▨	projection	behavior
signal		▨	▨	performance	▨	▨	▨	sensory impression	stimuli
context		▨	▨	▨	knowledge base	▨	▨	▨	ground
decoding		▨	▨	▨	▨	▨	▨	▨	interpretation
effect	info-as-knowledge	▨	work	essence	information object	intellectual entity	propositional content	intellectual entity	experience

2.2.1 *Message vs. meaning*

An object is a means by which its creator intends to communicate with a consumer. However, while an object can convey the creator's *message* – the numbers, words, images, sounds, etc. that constitute its information content – the *meaning* ascribed to that message is not actually carried by the object itself. Rather, the consumer's experience of cognitive meaning or emotional affect is an emergent epistemic effect of the consumptive act. An object mediating that act is a reflection of a particular mental state of its creator and is intended to induce a corresponding state on behalf of its consumer. However, since the consumer's mental formation of meaning arises through a contextually-grounded interpretation of the object's message, the creator's intention may never be fully realized [44]. While the potential for discordant interpretation may be minimal regarding the communication of propositional content, that is, content pertaining to objectively factual truth claims, individualistic responses are accepted and often even highly desirable outcomes for engagement with creative expressions.

2.2.2 *Ontological components*

In semiotic-theoretic terms, an act of object-mediated communication occurs when an expressible message is encoded into an object susceptible to contextualized decodings, resulting in subjectively experienced cognitive meanings or emotional affects (see Figure 1). In other words, an object reifies an abstract expressible thought, relative to some contextual frame of reference, into a consumable embodied thought, a critical distinction long established in the semiotic field, viz., *parole* vs. *langue*, or signifier vs. signified [25], as well as in library and information science, viz., work vs. document [39]. Following from this, the major ontological components of a digital object are its message, encoding of structure and form, carrier, behavior, and annotation, reflecting the chain of content reification (see Figure 2).

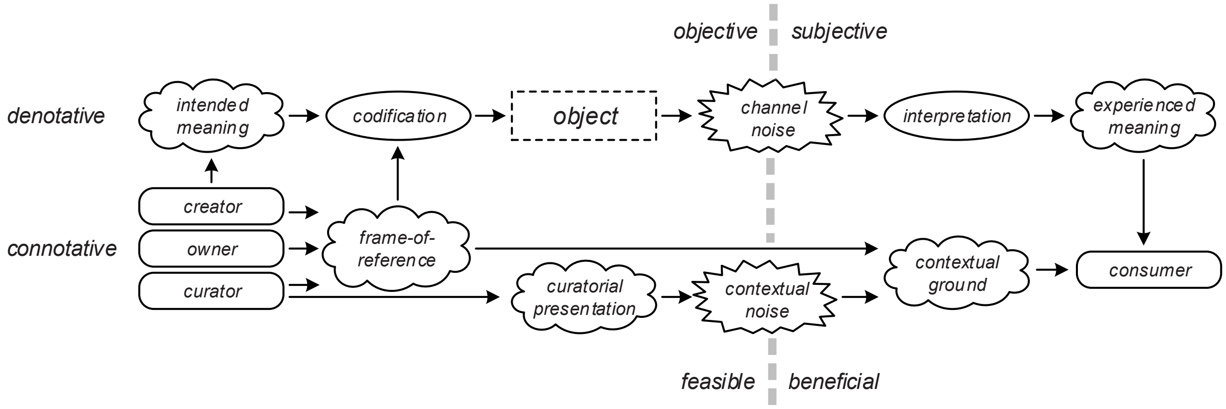


Figure 1. Object-mediated communication

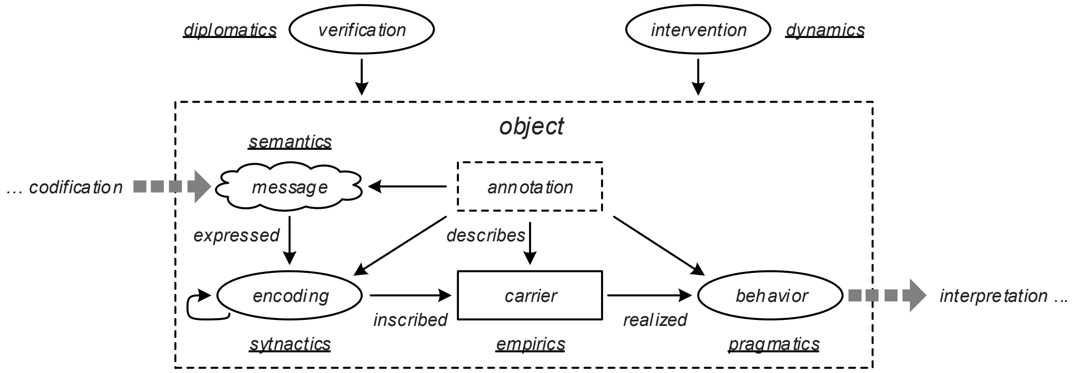


Figure 2. Object components

An object's message constitutes its semantic aspect, that is, the abstract information content it is intended to convey. This content is *expressed* through encodings into abstract symbol structures constituting the object's syntactic aspect [42] [14]. A given object may be distinguished by multiple hierarchically-nested encodings. These encodings can be distinguished between those concerned with the object's inner structure and outer morphological form. An object's symbolic expression is given tangible representation by being *inscribed* as a pattern of matter or energy on a physical carrier that constitutes the object's empiric aspect. This physical representation is made available for perception and interpretation by a consumer by being *realized* through behaviors that render the underlying information content in a human-sensible manner, constituting the object's pragmatic aspect.

The primary attribute of an object's inner structure is its format or type, which specifies the conventions of the object's symbolic expression and is the interface with its semantics [1]. The primary attribute of object morphology is identity. The identity of an object, like that of a linguistic sign, serves three purposes [13]:

1. As a *fence*, distinguishing and demarcating a particular object from all other potential objects;
2. As a *label*, facilitating unambiguous common reference to a singular object; and
3. As a *vehicle*, providing an actionable means for interacting with the object for some teleological purpose.

Morphological form also implicates the interface between the object and its empirics, that is, the encompassing computational environment necessary to support the object's visibility and dereferencing, for example, encoding details attendant to a file system, run-time environment, or network infrastructure depending upon whether the object is at rest or in motion. Without an assertion of identity, there is no effective way to establish or retrieve an object as the focus of curation scrutiny; similarly, without format typing, there is no effective means of interpreting and exploiting the object's message.

The attributes of identity and type are instances of annotations, propositional statements declaring specific characteristic values for significant object properties [12]. As these descriptive properties are fundamental to the successful interpretation and exploitation of an information object, they are a type of OAIS representation information [19]. (Representation information is also concerned with instrumental capabilities, for example, a viewer for a particular type; these are equivalent to Sept's pragmatic behaviors.) While identity and type are fundamentally *necessary* annotative properties, by themselves they may not be fully *sufficient* to ensure successful engagement, which may be dependent upon additional higher-order semantic and pragmatic properties [2]. Annotations provide the means to assert the perspectives or frames-of-reference of an object's creator, owner, and curatorial manager, and the interpretive experience of its consumer. The relationship between an annotation and its referent foreshadows that between the consumer's ground and interpretation: annotations contributing to the content's *objective* context and collectively informing the interpretive ground of the consumer's *subjective* experience.

2.2.3 Contextual ground

Traditional semiotic analysis presupposes two primary actors in the communicative act: creators and consumers. When dealing with curation of digital content, however, a third curatorial actor often intermediates between creator and consumer at the behest of a fourth, the content's owner. Content is traditionally collected, managed, and presented for use by a curatorial agent as part of larger aggregations based on explicit collection development policy, thematic unity, or administrative convenience. The contextual perspectives of the curatorial manager inevitably leave traces in a consumer's interpretive experience, just as a creator and owner's conceptual frames-of-reference inform the intention underlying content production. Thus, an object is inherently situated within a dynamic network of explicit and implicit denotative, connotative, and metaphorical associations by which it accumulates additional nuanced meanings or affects through the circumstances of its production, membership in curatorial aggregations, and under the imprimatur of its stewarding manager. Denotation refers to the overt commonplace meaning of an expression; connotation, to the indirect inferential meaning; and metaphor, to the allusive meaning [11]. These associations can take several forms:

1. Reputational assessments of individual and corporate content creators, owners, and curatorial stewards arising from a consumer's prior experience and professional judgment;
2. Intentions attendant to content production and ownership as expressed in collateral annotations;
3. Intentions attendant to content management as expressed through curatorial assessment, selection, arrangement, aggregation, and annotation;
4. Tangible relationships directly representable in content objects or object management systems, such as typed structural and semantic relationships between separate but dependent objects, and object aggregations and their subsidiary members; and
5. The tacit understanding acquired through experience or, in OAIS terms, as part of the knowledge base of a designated community that a consumer brings to the interpretive act [19].

All of these factors contribute to, but do not fully determine, the consumer's interpretive ground and subjective experience. The inherently recursive nature of these associational networks, in which every referent potentially can be the source of further references, is reminiscent of OAIS representation networks.

2.3 Object typology

Not every object will consist of the full complement of components. Thus, it is useful to distinguish between seven characteristic object types: blobs, artifacts, exemplars, products, assets, records, and heirlooms (see Table 2), which constitute a typology of increasingly specific definition and functional utility and value.

Table 2. Object typology

Differentia	Blob	Artifact	Exemplar	Product	Asset	Record	Heirloom
Dimension	empirics	syntactics	syntactics	semantics	pragmatics	diplomatics	dynamics
Mode	formative	informative	informative	informative	performative	evaluative	reformative
Act	inscription	identification	characterization	description	realization	verification	intervention
Concern	media	(outer) encoding	(inner) encoding	meaning / affect	experience	authenticity	persistence
Abstraction	carrier	form	structure	message	behavior	evidence	action
Quality	existential	intentional	purposeful	interpretable	useful	trustworthy	resilient
Value	nascent	incipient	potential	theoretical	practical	assured	enduring
Annotation	provenancial / administrative / permissive	morphological / relational / associational	structural	intellectual	instrumental	provenancial	provenancial

1. A blob is an *existential* object resulting from a formative act of inscription that produces tangible bits on an otherwise undifferentiated digital carrier, whether storage media or communication channel. Being opaque in all respects, nothing further can be known or inferred about a blob other than the fact of its existence. Thus, its value is nascent. Consider, for example, the bits …00000000110001101010010100… found somewhere on a carrier, which by themselves convey no recognizable, let alone useful, information.
2. An artifact is an *intentional* object resulting from an informative act of identification that demarcates a particular sequence of bits fixed in digital space-time. One can infer that an artifact was deliberately created, even if the purpose underlying the creation remains undisclosed. The essential properties of an artifact are its identity and symbolic encoding of outward-facing morphology. In and of itself, however, an artifact is syntactically opaque: it affords no opportunity to interpret or infer how its constituent bits express any underlying message. Thus, an artifact's value is incipient. Consider, for example, a named file with specific size,

timestamps, permissions, and MD5 digest, but absent any knowledge of its content's expression.

3. An exemplar is a *purposeful* object resulting from an informative act of characterization that documents the symbolic encodings of its internal structural expression. (The term "exemplar" is used in here in its non-qualitative sense of a general pattern or template without individuating characteristics.) The essential properties of an exemplar are its type or format and any further attributes entailed by that format. While these provide details of the exemplar's means of expression, its underlying message is still semantically opaque. Thus, an exemplar's value is potential. Consider, for example, a JPEG 2000-formatted image with three 8-bit components representing sRGB color samples, with 1024x1024 tiles, 64x64 code blocks, six decomposition layers, 25 quality layers, and 9-7 irreversible wavelet compression, but absent any knowledge of what the image represents.
4. A product is an *interpretable* object resulting from an informative act of description that documents its underlying message in terms appropriate to a particular domain of discourse. In and of itself, however, a product doesn't afford any practical means to experience or exploit that message. Thus, a product's value is theoretical. Consider, for example, the photographic image of Lake Merritt, a national historic landmark and the United States' first designated wildlife refuge located at 37.8039° N, 122.2591° W, close to UC3's offices in Oakland, California, absent any realizing behaviors.
5. An asset is a *useful* object resulting from a performative act of realization that exposes the product's message as stimuli apprehensible to human sensory modalities [5] [17]. Thus, an asset's value is practical: it can be directly experienced and exploited towards some useful purpose. Consider, for example, a consumer's experience engaging with the authentic Lake Merritt image in a colorimetric image processing environment supporting dynamic zooming, panning, cropping, annotation, etc., but absent any consideration of spatial or temporal extension.
6. A record is a *trustworthy* object resulting from an evaluative act of verification. The essential properties of a record are those important to considerations regarding the presumption, verification, and maintenance of authenticity and reliability [15]. Being trustworthy, a record's value is assured. Consider, for example, the Lake Merritt image that has been evaluated and determined to be what it purports to be, so that it can be accepted with confidence.
7. An heirloom is a *resilient* object resulting from a reformative act of proactive or reactive intervention that ensures the continuing viability and usability of the asset across space and time. Thus, to the extent to which those interventions are successful, an heirloom's value is enduring. Consider, for example, a consumer's future engagement experience with the Lake Merritt image.

The encodings underlying artifacts and exemplars may be hierarchically nested, for example, an artifact that is a file in a folder on a disk volume, or an exemplar that is a PCM sample stream inside a QuickTime multimedia wrapper inside of a Zip container.

The sequence of object types from blobs to heirlooms provides increasing functional utility and value, but the typology does not imply a strictly sequential inheritance hierarchy. While in practice many digital objects will have valid ontological identities across contiguous typological classes – for example, a product with known semantics, encoded in a known format (and thus, also an exemplar) and in well-characterized file (and thus, an artifact), inscribed on some tangible media (and thus, a blob) – this is not a necessary condition of the Sept model. Any higher-order type can effectively subclass directly from any inferior type. It is possible, for example, for product semantics to be known about an article whose inner encoding remains syntactically opaque. For example, consider an object about which the statement "This file is an image (of unknown format) of Lake Merritt" can be made. While one might have cause to question the accuracy of the assertion, it is nevertheless a valid case of a product being an artifact but not an exemplar. Similarly, it is possible for an asset to be an exemplar but not a product ("This JPEG image (of unknown subject) is viewable in that JPEG viewer"), a record to be a product but not an asset ("This image (with no format-specific viewer available) really *is* of Lake Merritt and *was* produced by the Lake Merritt Breakfast Club Foundation"); and an heirloom to be an asset but not a record ("This persistent viewable JPEG image *may* be of Lake Merritt").

2.3.1 Resilience

Resiliency ensures that an heirloom can be used for successful "communication with the future" [23] [22]. In information theory, factors that impede communication are considered noise [36]. In planning for effective interventions to ensure resiliency, the information-theoretic sender/receiver communication model distinguishes between *channel* and *contextual* noise: the former degrades the integrity of the signal, that is, the object carrying the encoded message, while the latter distorts the interpretive context of the object's message – for example, a conceptual misalignment between objective frame-of-reference and subjective contextual ground – and thus, the message's interpretation and ultimate effect on its receiver.

The primary strategy for ameliorating the effects of channel noise is the addition of redundancy to the encoded object, for example, mirroring, parity, checksums, erasure codes, etc. A strategy for minimizing contextual noise is to facilitate the most effective means for the creator, owner, and curatorial frames-of-reference to inform fully the contextual ground of the consumer; in other words, to ensure that the consumer can properly recover productive, proprietorial, and curatorial intentions. Descriptive annotations are included as a fundamental component of a digital object in order to facilitate this very process. However, since this strategy implies communication of the annotations across a channel either in conjunction with, or independent of, their referent content, the amelioration of contextual noise is itself subject to potential channel noise.

2.3.2 Annotation

Annotations are defined in terms of nine high-level categories: provenancial, administrative, relational, associational, permissive, morphological, structural, intellectual, and instrumental.

1. Provenancial annotations describe the actors, conditions, and events that led to the creation, acquisition, or revision of the content;
2. Administrative annotations describe the actors, conditions, and events related to the ongoing curation management of the content;
3. Relational annotations describe structural connections with other objects and aggregated collections.
4. Associational annotations describe frames-of-reference and curatorial policies and interpretive glosses.

5. Permissive annotations describe IPR and terms of service rights and obligations attendant to content management and engagement;
6. Morphological annotations describe content's externally-facing expression in terms of outer symbolic encodings;
7. Structural annotations describe content's internal expression in terms of inner symbolic encodings;
8. Intellectual annotations describe content in terms meaningful to an applicable domain of discourse; and
9. Instrumental annotations describe behaviors applicable to the content.

Table 2 indicates the earliest stage in the typological progression at which those particular annotation categories are relevant. For example, a blob has provenancial properties independent of and prior to any artifactual concerns (for example, carrier *A* was received from agent *B* at time *C*, etc.), an artifact has morphological properties independent of and prior to any exemplar-level concerns (file *X* of size *Y* and modification date *Z*, etc.), and so on.

2.4 Content engagement

Engagement with digital content is modeled in terms of four classes of actors and the lifecycle activities in which they participate [3] [8]. Content creators generate or acquire digital content and exercise *originating* intellectual and instrumental control and responsibility for the circumstances of that creation or acquisition; content owners exercise *ultimate* legal, financial, and permissive control and responsibility for its ongoing stewardship; content curators steward managed content and exercise *delegated* administrative, technical, and instrumental control and responsibility; and content consumers directly exploit or indirectly benefit from managed content for some individualistic purpose. The creator, curator, and consumer roles have a general correspondence to the producer, management, and consumer entities in the OAIS reference model [19]. The Sept consumer role, however, is more inclusive than its OAIS counterpart, encompassing any agent gaining some benefit from curated content through either direct *or* indirect means; while the Sept owner role and its concerns of proprietorial rights and obligations are not directly represented in the OAIS model. A given individual or corporate actor may hold these roles singly or a varying combinations at different points of time and in different organizational and operational contexts.

It is more useful to speak of the concerns of these roles in terms of an activity *continuum* rather than a lifecycle, as the latter implies a linear progression through clearly demarcated and distinguishable stages. In distinction, a continuum approach emphasizes the essential non-linear contiguity and overlapping interdependence of many curation activities and concerns [40]. Thus, it is more appropriate to group modes of engagement by thematic loci of concerns within a permeable continuum characterized by first-order catalyzation, concerned with creating, acquiring, or otherwise establishing resources of curation focus; second-order organization, concerned with codifying and imposing illuminating structure upon those resources; and third-order pluralization, concerned with expanding the reach and consequence of those resources. These continuum characteristics are based on the information continuum model (ICM) [34] although Sept's notion of catalyzation conflates the ICM's distinct creation and capture dimensions into a single category for purposes of conceptual parsimony.

While the thematic loci and continuum characteristics may seem synonymous – for example, production being equivalent to catalyze, etc. – they are actually orthogonal concerns: as indicated in Table 3, activities within each locus can be categorized by goals and intents spanning all three characteristic categories. Similarly, although terminological similarity implies a reductive association between roles and loci, for example, creators and production, etc., these are also orthogonal concerns, with each locus encompassing activities spanning each role. (Unfortunately, it is difficult to provide an intuitive depiction of the mutually-interdependent relationships of the three disparate dimensions of role, locus, and characteristic in a simple tabular form.) A comprehensive curation program will work towards promulgation of policies, strategies, and plans, and implementations of systems, services, operational procedures, and stakeholder guidance for all major continuum activities.

Table 3. Engagement continuum

Locus	Catalyze	Organize	Pluralize
Production	observe, simulate, create, derive	identify, classify, clean, annotate, package	license, submit, publish, cite, aggregate
Management	appraise, select, harvest, collect	normalize, characterize, arrange, annotate, store, index, plan, watch, intervene, administer	replicate, audit, notify, syndicate, resolve, authorize, report
Exploitation	search, discover, retrieve, subselect	analyze, correlate, synthesize, interpret, transform, annotate	summarize, validate, assert, refute

2.5 Policies and strategies

A formal statement of curation policy is necessary to set expectations properly and form the basis for acceptable terms of service and assessment of the efficacy of curation outcomes. Strategies represent specific organizational intentions for fulfilling or enforcing promulgated policies that can be implemented by concrete plans and activities [4] [37]. Curation policies, strategies, plans, and activities are modeled within Sept in terms of one preparatory and six implementation imperatives:

1. *Predilect*: decide what you intend;
2. *Collect*: obtain or effectuate what you decide;
3. *Protect*: preserve or sustain what you obtain;
4. *Introspect*: know what you protect;
5. *Project*: offer what you know; and
6. *Connect*: provide what you offer.
7. *Reflect*: (re)assess what you did.

While these imperatives are relevant to all aspects of the curation domain, for example, technical infrastructure, operational procedures, staffing, etc., they have the most obvious applicability to content. There is a general inheritance of relevant considerations across the imperative progression. The range of activities underlying these imperatives span the speculative and exploratory (that is, considerations of what *could* be), analytical and normative (what *should* be), operational (what *is*), and obligatory (what *must* be). In general, these activities should be proactive whenever possible, and reactive whenever necessary.

The foundational imperative for subsequent curation activities is *collection*, that is, bringing content into an appropriate stewardship environment under the control of a responsible curatorial manager with rights and obligations delegated from the content's owner. While it is *possible* that collected content may not be fully susceptible to successful curation outcomes, it is almost *certain* that uncollected content will be subject to curation failure with regard to current or future viability and availability [31]. The baseline level of curation assurance that can be realistically asserted by a responsible curating agent will generally be either as a blob or artifact, depending upon whether the content was collected as (undifferentiated) media or (opaque) files. Increasingly high-order outcomes may be possible if the content meets the incrementally more stringent criteria for exemplars, products, assets, records, and heirlooms.

Each imperative can be applied to every level of the typological hierarchy. While the resulting matrix (see Table 4) is suggestive of the NDSA levels of preservation [27], the typological progression plays a different role than the NDSA levels as it is defines increasing levels of general *utility* rather than specific *assurance*. However, concerns of assurance are encapsulated within the protect imperative. Thus, content utility and assurance both increase through the effective provisioning and implementation of progressive levels of environmental, administrative, technical, bibliographic, archival, access, and change control. Similarly, utility and assurance both increase through progressive levels of forensic, morphological, structural, intellectual, archival, and behavioral characterization arising from the introspect imperative.

Table 4. Policies and strategies

Imperative	Blob	Artifact	Exemplar	Product	Asset	Record	Heirloom
Predilect	service level agreement	disaster recovery / business continuity	format action plans	collection development policy	outreach and training	evidentiary standards	sustainability / succession planning
Collect	submission	packaging	normalization / canonicalization	discovery, workflow / tool integration	code / workflow repositories, aggregation	provenance	preservation planning tools
Protect	environmental control, media refresh, redundancy	administrative control, malware detection, fixity	technical control, migration	bibliographic control	access control, emulation	archival control	change control, preservation watch
Introspect	forensic characterization	morphological characterization, PID minting	structural characterization, ontologies, format registries	intellectual characterization, entity extraction, sentiment analysis	behavioral characterization, software registry	archival characterization, master registry	provenance, annotation
Project	media inventory	file inventory, PID resolution	object index	work catalog	transcoding, syndication, discovery	documentary form	versioned change history
Connect	legacy/emulated computational environments	file delivery	local format-aware processing	local disciplinary-specific processing	search/browse, hosted tools, annotation	authenticity-dependent workflows	consortial collaboration
Reflect	scrubbing	audit	tabletop testing	policy conformance	analytics	chain of custody	failure injection

3. CONCLUSION

The digital curation field has reached a stage of maturity where it can usefully draw upon a rich body of research and practical experience. Many specific segments of the curation domain have been subject to modeling activities, but the scope, coverage, and granularity of this work has varied widely. In an effort to ensure a comprehensive view of the domain for purposes of analysis, planning, and evaluation of its activities, the UC Curation Center has synthesized and reformulated the many valuable contributions of prior efforts into a new inclusive model. One important insight of the UC3 Sept modeling effort is that engagement with digital content is an inherently semiotic activity. Thus, the Sept model was developed by approaching all aspects of the curation domain through the lens of six characteristic dimensions: semantics, syntactics, empirics, pragmatics, diplomatics, and dynamics. The model conceives of a digital object as reifying abstract content into tangible form for purposes of mediated communication between a creator and consumer, carefully distinguishing between an object's message and meaning; the former being an objective embodiment of an expressed thought, while the latter is an emergent epistemic property arising from a subjective, contextualized reaction to the message. This leads to an object typology of progressively richer ontological basis and concomitant increasing content utility and value, consisting of blobs, artifacts, exemplars, products, assets, records, and heirlooms. Engagement with curated content is modeled by creator, owner, curator, and consumer agents and three loci of concerns for production, management, exploitation all operating within a continuum of originating, organizing, and pluralizing dimensions. Curation policies and strategies are modeled by seven imperatives: predilect, collect, protect, introspect, project, connect, and reflect.

The model components and its typology represent useful abstractions whose properties, coalescing around core conceptual centers of gravity, may be held by any particular component or typological instantiation. The components and typology can be used to make precise yet concise assertions regarding programmatic capabilities, intentions, actions, and outcomes. For example, it is common to divide preservation obligations into tripartite media, bit-level, and functional preservation levels. These correspond respectively to activities focused on ensuring the integrity of blobs, exemplars, and assets. Creating forensic disk images is a suitable strategy for preserving blobs (that is, media objects), independent of any artifactual morphology; fixity audit is a suitable strategy for artifacts (file objects), independent of any type characterization; migration, for products (syntactically- and semantically-characterized objects), independent of any behavioral considerations; and emulation, for assets (experiential objects).

While a curating agent could choose to enforce a lower service obligation than what may be otherwise supportable by an object's typological characteristics, it is not possible to meet a higher obligation. For example, a digital exemplar (that is, a *typed* file) could be managed purely as an artifact (an *opaque* file) through the expedient of disregarding any non-morphological characterization, but no matter how successful the preservation of a true artifact, it will never afford any higher-order structural information about its contents; if such information were known or could be inferred, the object would be an exemplar rather than an artifact. Thus, finely-grained typological modeling permits more precise statements of curation intention, expectation, and result. For example, saying that an object will be "functionally" preserved is open to potential ambiguity; on the other hand, saying that it will be preserved as an exemplar makes clear that it will continue to be a purposeful object through persistent association with pertinent inner structural encoding information. Similarly, a preserved product will remain interpretable through association with appropriate semantic characterization, and a preserved asset will remain useful through association with realizing behaviors.

Given a semiotic view of content engagement, it may never be possible to preserve a digital object "perfectly." While it is potentially possible to fix and maintain indefinitely the state for components on the objective side of the communication divide, i.e., message, encoding, carrier, annotation, and behavior (see Figure 1), on the subjective side, the consumer's future contextual ground is not susceptible to any equivalent constraint as it is contingent on the totality of that consumer's intervening lived experience. This may not be significant for propositional content consisting of purportedly-objective factual claims, but could be important for creative content.

All of the Sept model components were developed incrementally from first principles in an effort to ensure comprehensive applicability and internal consistency. The use of such a model is important for increasing confidence that programmatic planning is systematic and not ad hoc. While the model introduces unfamiliar terminology, UC3 believes that this vocabulary supports important nuanced distinctions in the delineation of content, content engagement, and curation policies and strategies. The Sept model's granular definition permits the concise statement of common curation intentions, activities, and outcomes. It forms the basis for UC3's decision-making processes regarding curation infrastructure, services, and initiatives, and may be of interest to the wider curation community, with which it shares many common concerns and practices.

4. ACKNOWLEDGMENTS

I wish to thank Christoph Becker, Michael Buckland, Jay Gattuso, Asen Ivanov, John Kunze, Emily Maemura, Jerry McDonough, Nathan Moles, Sheila Morrissey, Seamus Ross, Barbara Sierman, Nicholas Taylor, and the anonymous iPRES reviewers for their many perceptive comments that have materially improved the Sept domain model and its exposition in this paper.

5. REFERENCES

[1] Abrams, S. 2007. File formats. In *Curation Reference Manual*, Digital Curation Centre, URL=http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/file-formats.

[2] APARSEN. 2014. *D11.3 report on a common vision of digital preservation: Progress to year 3*. Technical report. URL= http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2014/06/APARSEN-REP-D11_3-01-1_1_inclURN.pdf.

[3] Ball, A. 2012. *Review of Data Management Lifecycle Models*. University of Bath. URL=http://opus.bath.ac.uk/28587/1/redm1rep120110ab10.pdf.

[4] Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., and Hoffman, H. 2009. Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. *Intl. J. Dig. Lib.* 10, 4 (December), 133-157. DOI=doi:10.1007/s00799-009-0057-1.

[5] Benyon-Davis, P. 2011. *Significance: Exploring the Nature of Information, Systems and Technology*. Palgrave Macmillan, New York.

[6] Berlo, D. 1960. *The Process of Communication*. Holt, Rinehart, and Winston, New York.

[7] Buckland, M. 1991. Information as thing. *J. Am. Soc. Inform. Sci.* 42, 5 (June), 351-360.

[8] CEOS Working Group on Information Systems and Services. 2012. *Data Life Cycle Models and Concepts.* URL=http://www.ceos.org/images/DSIG/Documents/Data_Lifecycle_Models_and_Concepts_v13-1.docx.

[9] Cheney, J., Lagoze, C., and Botticelli, P. 2001. Towards a theory of information preservation. In *5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL '01)* (Darmstadt, September 4-9). HDL=http://hdl.handle.net/1853/5828.

[10] Dallas, C. 2007. An agency-oriented approach to digital curation theory and practice. In *ICHIM ''07, International Cultural Heritage Informatics Meeting* (Toronto, October 24-26), 49-72. URL=http://www.archimuse.com/ichim07/papers/dallas/dallas.html.

[11] Danesi, M. 2003. Metaphorical 'networks' and verbal communication: A semiotic perspective of human discourse. Σημειωτκή – *Sign Systems Studies* 2, 341-364. URL=http://www.ceeol.com/aspx/getdocument.aspx?id=37f0a98fa71241c790a6a1d1af38db9c.

[12] Dappert, A., and Farquhar, A. 2009. Significance is in the eye of the beholder. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries* (Corfu, September 27-October 2, 2009). URL=http://www.planets-project.eu/docs/papers/Dappert_SignificantCharacteristics_ECDL2009.pdf.

[13] Dewey, J. 1910. *How We Think*. D. C. Heath, Boston.

[14] Doerr, M., and Tzitzikas, Y. 2012, *Information Carriers and Identification of Information Objects: An Ontological Approach*. Technical report. URL=http://arxiv.org/abs/1201.0385.

[15] Duranti, L., ed. 2005. *The Long-Term Preservation of Authentic Electronic Records: Findings of the InterPARES Project*. Archilab, San Miniato. URL= http://www.interpares.org/book/index.htm.

[16] Flouris, G., and Maghini, C. 2007. Terminology and wish list for a formal theory of preservation. In *PV 2007 – Ensuring the Long-Term Preservation and Value Adding to Scientific and Technical Data* (Munich, October 9-11). URL=http://www.pv2007.dlr.de/Papers/Flouris_WishListPreservation.pdf.

[17] Heslop, H., Davis, S., and Wilson, A. 2002. *An Approach to the Preservation of Digital Records*. National Archives of Australia. URL=http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm16-47161.pdf.

[18] IFLA Study Group on the Functional Requirements for Bibliographic Records. 1998. *Functional Requirements for Bibliographic Records: Final Report*. K. G. Saur, München. URL=http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf.

[19] ISO 14721. 2012. *Space data and information systems – Open archival information system (OAIS) – Reference model*. URL= http://public.ccsds.org/publications/archive/650x0m2.pdf.

[20] Kahn, R., and Wilensky, R. 1995. A framework for distributed digital object services. *Intl. J. Digital Libraries* 6, 2 (April), 115-123. DOI=http://dx.doi.org/10.1007/s00799-005-0128-x.

[21] Liebenau, J., and Backhouse, J. 1990. *Understanding Information: An Introduction*. Macmillan, London.

[22] Mois, M., Klas, C.-P., and Hemmje, M. L. 2009. Digital preservation as communication with the future. In *16th International Conference on Digital Signal Processing* (Santorini, July 5-7). DOI=http://dx.doi.org/10.1109/ICDSP.2009.5201104.

[23] Moore, R. 2008. Towards a theory of digital preservation. *Intl. J. Digital Curation* 3, 1 (June), 63-75. DOI=http://dx.doi.org/10.2218/ijdc.v3i1.42.

[24] Morris, C. 1946. *Signs, Language, and Behavior*. Prentice-Hall, New York.

[25] Nöth, W. 1990. *Handbook of Semiotics*. Indiana University Press, Bloomington.

[26] Peirce, C. S. 1932. *Collected Papers of Charles Sanders Peirce. Volumes I and II: Principles of Philosophy and Elements of Logic*. Harvard University Press, Cambridge.

[27] Philips, M, Bailey, J., Goethals, A., and Owens, T. 2013. *The NDSA Levels of Digital Preservation: An Explanation and Uses*. Library of Congress. URL=http://www.digitalpreservation.gov/ndsa/working_groups/docuemnts/NDSA_Levels_Archiving_2013.pdf.

[28] PREMIS Editorial Committee. 2012. *PREMIS Data Dictionary for Preservation Metadata*. URL=http://www.loc.gov/standards/v2/premis-2-2.pdf.

[29] Ranganathan, S. R. 1931. *The Five Laws of Library Science*. Madras Library Association.

[30] Reynolds, P. D. 1971. *A Primer in Theory Construction*. Bobbs-Merrill, New York.

[31] Rosenthal, D. S. H. 2014. Improving the odds of preservation. *CNI Fall 2014 Membership Meeting* (Washington, December 8-9). URL=http://www.cni.org/topics/digital-prservation/improving-the-odds-of-preservation.

[32] Ross, S. 2007. Digital preservation, archival science, and methodological foundations for digital libraries. In *ECDL 2007, The 11th European Conference on Digital Libraries* (Budapest, September 16-21). URL=http://www.ecdl2007.org/Keynote_ECDL2007_SROSS.pdf.

[33] Rumsey, A. S. 2010. Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. URL=http://blueribbontaskforce.sdsc.edu/biblio/BRTF_Final_Report.pdf.

[34] Schauder, D., Johanson, G., and Stillman, L. 2005. Sustaining a community network: The information continuum, and the case of VICNET. *J. Community Informatics* 1, 2. URL=http://ci-journal.net/index.php/ciej/article/view/239/203.

[35] Schram, W. 1954. How communication works. In *The Process and Effects of Communication*. University of Illinois Press, Urbana.

[36] Shannon, C., and Weaver, W. 1949. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana-Champaign.

[37] Sierman, B. 2014. The SCAPE policy framework, maturity levels, and the need for realistic preservation policy. In *Proceedings of the 11th International Conference on Digital Preservation* (Melbourne, October 6-10), 259-266. URL= https://www.nla.gov.au/sites/default/files/ipres2014-proceedings-version_1.pdf.

[38] Stamper, R. 1973. *Information in Business and Administrative Systems*. Wiley, New York.

[39] Svenonius, E. 2000. *The Intellectual Foundations of Information Organization*. MIT Press, Cambridge.

[40] Upward, F. 1996. Structuring the records continuum – Part one: Postcustodial principles and properties. *Archives and Manuscripts* 24, 2 (November), 268-285. URL= http://www.infotech.monash.edu.au/research/groups/rcrg/publications/recordscontinuum-fupp1.html.

[41] UC Curation Center. 2015. *Digital Curation Foundations*. URL=http://wiki.ucop.edu/display/Curation/Foundations.

[42] Waters, D., and Garret, J. 1996. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*. URL=http://www.clir.org/pubs/reports/pub63.

[43] Wickett, K., Sacchi, S., Durbin, D., and Renear, A. 2012. Identifying content and levels of representation in scientific data. *Proceedings of the ASIS&T 75th Annual Meeting* (Baltimore, October 26-31). URL= https://www.ideals.illinois.edu/handle/2142/35259.

[44] Wimsatt, W., and Beardsley, M. 1946. The intentional fallacy. *Sewanee Review* 54, 468-4.

Participatory Digital Repositories for the Curation of Performing Arts with Digital Technology

Guillaume Boutard
University at Buffalo
Department of Library and Information Studies
534 Baldy Hall
Buffalo, NY 14260-1020
gboutard@buffalo.edu

ABSTRACT

The complexity of socio-technical systems in artistic production involving digital technology, especially in the performing arts, challenges digital curation models with a potential shift from cycles to networks. We argue that digital curation models need to develop in parallel to interdisciplinary investigations of these systems. These investigations question the conceptual separation of curation stages as well as roles. In this paper, we build on previous curation projects for new media arts and on the historical analysis of a specific work of contemporary music with live electronics to propose future directions for the integration of curation practices, artistic practices and digital curation models.

General Terms

Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice.

Keywords

Digital curation; Artistic practice; Music with live electronics; Participatory digital repositories

1. INTRODUCTION

Abbott [1] emphasizes the relevance of digital curation models in the artistic domain, especially in the domain of performing arts, where the goal is to provide the means for new interpretations. The development of lifecycles in relation to artistic production has long been documented as well as collaborative properties of production processes [3]. From an organizational point of view, Benghozi [4] described the artistic production context as *ad hoc* and building on ephemeral organizations involving flexible collaborations and strong commitment of the agents.

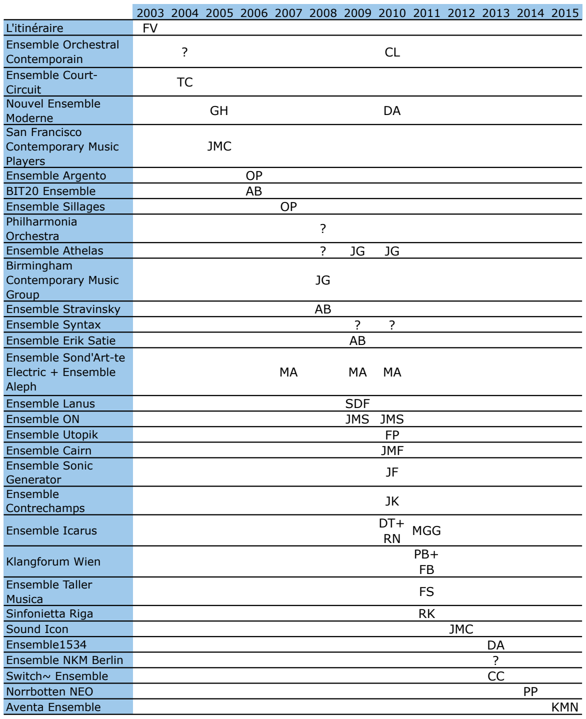
While institutional repositories, in relation to research activities, have developed since the early twenty-first century [17], Molloy [19] argues that in the performing arts domain: "the motivation and the enthusiasm for good digital curation practice are both present; awareness, training and reward structures for improved digital curation practice are currently absent" (p. 19). The situation is similar in the domain of contemporary music involving live digital technologies, despite several pioneering projects such as Mustica [5] at Institut de Recherche et Coordination Acoustique/Musique (IRCAM). One reason might be the inadequacy of curation lifecycle implementations with regard to work practices, involving ephemeral organizations but strong commitment.

2. CREATIVE PROCESSES AND LIFECYCLES

Creative processes have gained research attention in various disciplines in relation to diverse artistic domains. Prior [23], investigating experimental practices in avant-garde electronic music from an actor-network theory perspective, states that "[...] it is certainly not the case in music production that sociological questions are more relevant at the point at which the product finds its way through distribution processes, leaving the creative process itself to aesthetics or musicology" (p. 315).

Generally speaking, in a work community, work practice involves repetition and adaptation. Nathanael and Marmaras [20] describe practice adaptation with a situated action and cognition angle: "practice adaptations will typically involve both the minds and bodies of people participating in the community as well as their tools and other material arrangements" (p. 365). In the contemporary music context, Donin and Theureau [10, 11] discuss the temporal aspects of compositional processes in relation to the development of a body of work. They base their arguments, notably, on the study of the work of composer Philippe Leroux and the relations between several pieces, specifically, Voi(rex) and Apocalypsis. They conceptualize the notion of situated composition, in which the tools are critical: "[...] the content and organization of the composer's studio (computer and software included) is a relative invariant built up over a number of years. Long timespan creative cognition, unique individual cognition and situated cognition appear as constituting three related characteristics. In this way, we may speak of a unique individual cognition of a technically situated actor" (p. 247).

Furthermore, from a social perspective on the domain of contemporary music with live electronics, the significance of computer music designers in the creative process has been emphasized in the literature [27]. This situation tends to increase the complexity of the social context of production.



	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
L'itinéraire	FV												
Ensemble Orchestral Contemporain		?						CL					
Ensemble Court-Circuit		TC											
Nouvel Ensemble Moderne			GH					DA					
San Francisco Contemporary Music Players			JMC										
Ensemble Argento				OP									
BIT20 Ensemble				AB									
Ensemble Sillages					OP								
Philharmonia Orchestra						?							
Ensemble Athelas						?	JG	JG					
Birmingham Contemporary Music Group						JG							
Ensemble Stravinsky						AB							
Ensemble Syntax							?	?					
Ensemble Erik Satie							AB						
Ensemble Sond'Art-te Electric + Ensemble Aleph					MA		MA	MA					
Ensemble Lanus							SDF						
Ensemble ON							JMS	JMS					
Ensemble Utopik								FP					
Ensemble Cairn								JMF					
Ensemble Sonic Generator								JF					
Ensemble Contrechamps								JK					
Ensemble Icarus								DT+ RN	MGG				
Klangforum Wien									PB+ FB				
Ensemble Taller Musica									FS				
Sinfonietta Riga									RK				
Sound Icon										JMC			
Ensemble1534											DA		
Ensemble NKM Berlin											?		
Switch~ Ensemble											CC		
Norrbotten NEO												PP	
Aventa Ensemble													KMN

Figure 1: Live electronic musicians (only their initials are provided) involved in the performances of Voi(Rex) from 2003 to 2015 (adapted from Plessas and Boutard, 2015).

Plessas and Boutard [22] distinguish between agents involved in the production of the live software and live performers of the electronic part of the work as those who interpret *the* and *with the* software. They base their investigation of the work of live electronics musicians, notably, on a historical review of the performances of a specific work: Voi(rex) by composer Philippe Leroux. Their case study reveals that the activity of live electronics performance and interpretation could benefit from a complex network of expertise, distributed and developed across time and space, rather than a system defined as a cycle of subsequent improvements. While the electronic part of the work was migrated several times in relation to technological obsolescence, the distribution of human agents in charge of the performance follows a different scheme (see Figure 1). Some of them performed the work several times during the same year, with the same version of the software, but with different ensembles in multiple locations (i.e. different production contexts). Some of them have performed the work at multiple stages of its technological development, sometimes several years after the first performance.

Critically, the modification of the live software is not just related to technological obsolescence but also to the very process of interpretation. Each production of the work is situated in a specific human and technological context (involving performing spaces) and requires a process of interpretation. A parallel view in visual arts is presented by Dekker [9]:

> Important to note in this respect are observations by people who have conducted case studies that it is easier to document a work when it is presented. When a work is in storage it is much harder to talk about specific issues. The installation of a work facilitates the detection of problems and provides a better view on the specific decisions taken or methods used in the creation of the work. It is for this reason that some people argue for more presentations to enhance the visibility and understanding of the way art works (Dekker, 2010). It could be argued that presentation leads to preservation. (p. 163)

From this perspective, use, dissemination and preservation actions collapse into one activity, which, ironically, is distributed across time and space. Plessas and Boutard [22] argue that 1) the non-linearity of the multiple aspects of the production of musical works with live electronics and 2) the absence of a clear separation between bugs and interpretation, question the ways we collect expertise and manage the electronic part of the work over time. We argue that this situation questions the way we curate these works, from the perspective of digital curation lifecycles.

3. PARTICIPATORY DIGITAL REPOSITORIES

Rinehard and Ippolito [25] describe four preservation strategies: storage, emulation, migration, and re-interpretation. Notoriously, Rinehard is a strong advocate for re-interpretation, a strategy which relates notably to the notion of variable media and the Media Art Notation System [24].

> A reinterpretation sacrifices basic aspects of the work's appearance in order to retain the original spirit. Rare for the fine arts, reinterpretation is common in dance and theater, although even in the performance arts its use can be controversial. [25, p. 10]

Reinterpretation is closely related to the notion of performance, and though it is a controversial view, they argue that "[...] society has to move from preserving media to preserving art. In the process, we will have to view change not as an obstacle but as the means of survival" (p. 46). The performance requires re-interpretation of the socio-technical framework with a constant investigation of the boundary between migration and interpretation [22].

According to Caplan, Kehoe and Pawletko [7], "there is wide agreement in the international preservation community that responsibility for long-term preservation of scientific and cultural heritage materials must be shared among many organizations" (p. 35). The distributed property of long-term preservation is not limited to the organizational level. Kunda and Anderson-Wilk [17] state that "[...] digital preservation is only one aspect of the larger, necessarily collaborative role of digital curation" (p. 896). Kaufmann [15] provides an example, in the artistic domain, of distributed expertise at the individual level (the use of forums of expertise for the preservation of artworks in relation to specific digital technologies).

In this context the question of stakeholders in digital curation is critical. Dappert and Farquhar [8], state that "in the digital preservation context, significance is determined by the stakeholders involved in the preservation process. These include the producer of the digital object, the custodian who holds it, and the consumer who will access it" (p. 302). The sociology of art has brought into light the role of intermediary professions in relation to art production, especially in relation to technology [18]. Konstantelos [16] argues: "viewing software art as a sociotechnical system – where the development of artwork binds people, processes and technology in a joint and collaborative effort – could lead to a (re-)appraisal of our understanding of context" (p. 18-19). Similarly, in the new media arts domain, Obermann [21] proposes to include assistants in the documentation process. On the other hand, creative processes are unique and attempts at modeling roles and interactions, e.g. the Capturing Unstable Media Conceptual Model (CMCM) developed by V2_Organisation [12], have strong shortcomings: "notwithstanding the high value of their theoretical underpinnings, one of the pitfalls of all the models discussed, especially those of VMQ, MANS, and CMCM, is their highly prescribed structures which, as said before, makes it difficult to implement a realistic and easily repeatable documentation project in conservation practice, especially outside the field of installation art" [9, p. 164]. Consistently with their proposition for re-interpretation, Rinehardt and Ippolito [25] go a step further; they "[...] reject the notion that a bunch of preservation experts in a room will someday concoct a one-size-fits-all technical fix to rescue culture from oblivion. Instead, we see rescuing new media as a task that is best distributed across a wide swath of cultural producers and consumers, who will choose the most appropriate strategy for each endangered work, one by one" (p. 10). Rinehardt and Ippolito's statement leads to the discussion about convergence between crowdsourcing and preservation: "this potential for crowdsourcing the preservation of context is one reason that the Variable Media Questionnaire now encourages input on artwork's essence not just from the creators and curators close to a project, but from those with no more claim to authority than the average gallery-goer" (p. 178).

In light of the Voi(rex) case study, the socio-technical system, emphasized by Konstantelos [16], is a complex network of human experience and technological migration and (re-)interpretation (i.e. adaptation of the software to the current production context of the work as described by Plessas and Boutard [22]). The goal to integrate this situation at the curation level has three consequences:

1. the need for collaboration repositories, in the sense given by Treloar, Groenewegen and Harboe-Ree [26], that is to say, as opposed to publication repositories;
2. the need for non-linear curation systems that fit practices; and
3. the need to propose non-prescriptive (i.e., not based on formal models constraining the definition of the creative process) documentation methodologies.

Several initiatives in the artistic domain have built on crowdsourcing and distributed expertise, such as Rhizome and the Archive of Digital Art (ADA). Authors have emphasized the use of new technologies for curation purposes. For example, Kunda and Anderson-Wilk [17] state that "in the last several years, with the rise of Web 2.0 and social computing, our institutions of record are facing a new digital curation challenge: stakeholder communities of interest are now expecting customized Web interfaces to the institutional knowledge repositories, online environments where community members can contribute content and see themselves represented, as well as access the archived resources" (p. 896). In the context of moving image archives, Gracy [13] states: "in some ways, it is inevitable that social networks should extend into the work of cultural institutions, as they have infiltrated other institutions (such as education and government)" (p. 185).

The question is then: which methodological framework for curation fits the need for participatory digital repositories? The Digital Curation Centre's (DCC) curation lifecycle [14] is linear within its circularity; it builds on the Open Archival Information System's (OAIS) input-output/producer-consumer model by connecting both ends with a focus on use and re-use. It lacks potential for integrating lessons learned from these 'last several years' as well as recent propositions based, notably, on interactionism and activity theory [6].

In light of the Voi(rex) case study, new approaches to digital curation require participation and interaction at every so-called stage of the curation lifecycle, creating a complex network of interactions among all the stakeholders. Barry, Born and Weszkalnys [2] describe three modes of interdisciplinarity: 1) integrative-synthesis; 2) subordination-service; and 3) agonistic-antagonistic, where "[...] interdisciplinary research is conceived neither as a synthesis nor in terms of a disciplinary division of labour, but as driven by an agonistic or antagonistic relation to existing forms of disciplinary knowledge and practice. Here, interdisciplinarity springs from a self-conscious dialogue with, criticism of or opposition to the intellectual, ethical or political limits of established disciplines or the status of academic research in general [...]" (p. 29). They further describe three rationals motivating interdisciplinary research: 1) accountability, "[...] breaking down the barriers between science and society [...]" (p.31) ; 2) innovation ; and 3) ontology, questioning models, assumptions and values. The logic of ontology is thus a driving force for a truly agonistic-antagonistic interdisciplinary research. The need for re-envisioning the question of curation lifecycle, stakeholders and creative processes is fundamentally an interdisciplinary question, which requires an agonistic-antogonistic approach.

4. CONCLUSION

The study of artistic practices involving digital technologies, especially in the performing arts, tends to put into a different light the vision of digital curation as a simple lifecycle. The assumption that "the use and interaction that takes place between the community and the digital resources, within the curated Web space, is the breeding ground for new, improved formulations of knowledge, which are then deposited into the IR [Institutional repository]" [17, p. 905], requires to posit an *a priori* conceptual boundary between knowledge production and digital repositories. This position leaves the repository outside of what Rinehardt and Ippolito refer to as the 'essence' of the work. There is an urgent need to question this boundary and, as a consequence, the roles (and the range) of the stakeholders.

The theoretical framework for new models of curation requires interdisciplinary research including: digital curation; computer supported cooperative work, building on its ethnomethodological roots [2]; activity theory and work psychology [6]. The strong commitment of agents in the artistic domain, as described by Benghozi [4], supports the possibility as well as the necessity to do so.

5. REFERENCES

[1] D. Abbott. Preserving Interaction. In L. Konstantelos, J. Delve, D. Anderson, C. Billenness, D. Baker, and M. Dobreva, editors, *Software Art*, volume 2 of *The Preservation of Complex Objects*, pages 61–70. JISC, 2012.

[2] A. Barry, G. Born, and G. Weszkalnys. Logics of interdisciplinarity. *Economy and Society*, 37(1):20–49, Jan. 2008.

[3] H. S. Becker. Art As Collective Action. *American Sociological Review*, 39(6):767, Dec. 1974.

[4] P.-J. Benghozi. Les sentiers de la gloire: savoir gérer pour savoir créer. In F. Charue-Duboc, editor, *Des savoirs en action*, pages 51–87. L'Harmattan, Paris, France, 1995.

[5] A. Bonardi, J. Barthélémy, G. Boutard, and R. Ciavarella. First Steps in Research and Development about the Sustainability of Software Modules for Performing Arts. In *proceedings of Journées d'Informatique Musicale JIM'08*, Albi, 2008.

[6] G. Boutard and F. Marandola. Mixed music creative process documentation methodology: outcomes of the DiP-CoRE project. In *proceedings of the 9th Conference on Interdisciplinary Musicology*, Berlin, 2014.

[7] P. Caplan, W. R. Kehoe, and J. Pawletko. Towards Interoperable Preservation Repositories: TIPR. *International Journal of Digital Curation*, 5(1):34–45, June 2010.

[8] A. Dappert. Metadata for Preserving Computing Environments. In D. Anderson and J. Delve, editors, *Gaming Environments and Virtual Worlds*, volume 3 of *The Preservation of Complex Objects*, pages 63–73. JISC, 2013.

[9] A. Dekker. Enjoying the gap: Comparing contemporary documentation strategies. In J. Noordegraaf, C. G. Saba, B. Le Maître, and V. Hediger, editors, *Preserving and Exhibiting Media Art: Challenges and Perspectives*, pages 150–169. Amsterdam University Press, 2013.

[10] N. Donin and J. Theureau. Theoretical and methodological issues related to long term creative cognition: the case of musical composition. *Cognition, Technology & Work*, 9(4):233–251, Oct. 2007.

[11] N. Donin and J. Theureau. La coproduction des oeuvres et de l'atelier par le compositeur (À partir d'une étude de l'activité créatrice de Philippe Leroux entre 2001 et 2006). *Circuit : musiques contemporaines*, 18(1):59–71, 2008.

[12] S. Fauconnier and R. Frommé. Capturing Unstable Media. In *Proceedings of the 8th International Cultural Heritage Informatics Meeting (ICHIM 04)*, Berlin, Germany, 2004.

[13] K. F. Gracy. Moving Image Preservation and Cultural Capital. *Library Trends*, 56(1):183–197, 2007.

[14] S. Higgins. The DCC Curation Lifecycle Model. *The International Journal of Digital Curation*, 3(1):134–140, Dec. 2008.

[15] F. Kaufmann. Hacking Mondrian. In B. Serexhe, editor, *Digital Art Conservation: Theory and practice*, pages 273–284. Ambra, Karlsruhe, 2013.

[16] L. Konstantelos. Documenting the Context of Software Artworks through Social Theory: Towards a Vocabulary for Context Classification. In L. Konstantelos, J. Delve, D. Anderson, C. Billenness, D. Baker, and M. Dobreva, editors, *Software Art*, volume 2 of *The Preservation of Complex Objects*, pages 18–32. JISC, 2012.

[17] S. Kunda and M. Anderson-Wilk. Community Stories and Institutional Stewardship: Digital Curation's Dual Roles of Story Creation. *Libraries and the Academy*, 11:895–914, 2011.

[18] P.-M. Menger and D. Cullinane. Technological Innovations in Contemporary Music. *Journal of the Royal Musical Association*, 114(1):92–101, Jan. 1989.

[19] L. Molloy. Digital curation skills in the performing arts – an investigation of practitioner awareness and knowledge of digital object management and preservation. *International Journal of Performance Arts and Digital Media*, 10(1):7–20, 2014.

[20] D. Nathanael and N. Marmaras. On the development of work practices: a constructivist model. *Theoretical Issues in Ergonomics Science*, 9(5):359–382, Aug. 2008.

[21] A. Obermann. Digital Media Art Demands Commitment! In *Digital Art Conservation: Theory and practice*, pages 315–322. Ambra, Karlsruhe, 2013.

[22] P. Plessas and G. Boutard. Transmission et interprétation de l'instrument électronique composé. In *Proceedings of JIM2015*, Montreal, QC, May 2015.

[23] N. Prior. Putting a Glitch in the Field: Bourdieu, Actor Network Theory and Contemporary Music. *Cultural Sociology*, 2(3):301–319, Nov. 2008.

[24] R. Rinehart. A System of Formal Notation for Scoring Works of Digital and Variable Media Art, June 2004.

[25] R. Rinehart and J. Ippolito. *Re-collection: Art, New Media, and Social Memory*. The MIT Press, Cambridge, MA, June 2014.

[26] A. Treloar, D. Groenewegen, and C. Harboe-Ree. The data curation continuum: managing data objects in institutional repositories. *D-Lib Magazine*, 13(9-10), 2007.

[27] L. Zattra. Les origines du nom de RIM (Réalisateur en informatique musicale). In *proceedings of the JIM* 2013, Saint-Denis, 2013.

Archiving Deferred Representations Using a Two-Tiered Crawling Approach

Justin F. Brunelle, Michele C. Weigle, and Michael L. Nelson
Old Dominion University
Department of Computer Science
Norfolk, Virginia, 23508
{jbrunelle, mweigle, mln}@cs.odu.edu

ABSTRACT

Web resources are increasingly interactive, resulting in resources that are increasingly difficult to archive. The archival difficulty is based on the use of client-side technologies (e.g., JavaScript) to change the client-side state of a representation after it has initially loaded. We refer to these representations as *deferred representations*. We can better archive deferred representations using tools like headless browsing clients. We use 10,000 seed Universal Resource Identifiers (URIs) to explore the impact of including PhantomJS – a headless browsing tool – into the crawling process by comparing the performance of wget (the baseline), PhantomJS, and Heritrix. Heritrix crawled 2.065 URIs per second, 12.15 times faster than PhantomJS and 2.4 times faster than wget. However, PhantomJS discovered 531,484 URIs, 1.75 times more than Heritrix and 4.11 times more than wget. To take advantage of the performance benefits of Heritrix and the URI discovery of PhantomJS, we recommend a tiered crawling strategy in which a classifier predicts whether a representation will be deferred or not, and only resources with deferred representations are crawled with PhantomJS while resources without deferred representations are crawled with Heritrix. We show that this approach is 5.2 times faster than using only PhantomJS and creates a frontier (set of URIs to be crawled) 1.8 times larger than using only Heritrix.

Categories and Subject Descriptors

H.3.7 [**Online Information Services**]: Digital Libraries

General Terms

Design, Experimentation, Measurement

Keywords

Web Architecture, HTTP, Web Archiving, Memento

1. INTRODUCTION

The Web – by design and demand – continues to change. Today's Web users expect Web resources to provide application-like interactive features, client-side state changes, and personalized representations. These features enhance the browsing experience, but make archiving the resulting representations difficult – if not impossible. We refer to the ease of archiving a Web resource as *archivability* [8].

Web resources are ephemeral by nature, making archives like the Internet Archive [24, 36] valuable to Web users seeking to revisit prior versions of the Web. Users (and robots) utilize archives in a variety of ways [3, 15, 18]. Live Web resources are more heavily leveraging JavaScript (i.e., Ajax) to load embedded resources, which leads to the live Web "leaking" into the archive [9] or missing embedded resources in the archives, both of which ultimately results in reduced archival quality [7].

We define *deferred representations* as those representations of resources that rely on JavaScript and other client-side technologies to load embedded resources after the initial page load. We use the term *deferred* because the representation is not fully realized and constructed until *after* the JavaScript code is executed on the client. Conventional Web crawlers (e.g., Heritrix, wget) are not equipped with the necessary tools to execute JavaScript during the archival process [6] and subsequently never dereference the URIs of the resources embedded via JavaScript and are required to complete the deferred representation. PhantomJS allows JavaScript to execute on the client, rendering the representation as would a Web browser. In the archives, the missing embedded resources return a non-200 HTTP status (e.g., 404, 503) when their Universal Resource Identifiers (URIs) are dereferenced, leaving pages *incomplete*. Deferred representations can also lead to *zombies* which occur when archived versions of pages inappropriately load embedded resources from the live Web, leaving pages incorrect, or more accurately, *prima facie violative* [2].

We investigate the impact of crawling deferred representations as the first step in an improved archival framework that can replay deferred representations both completely and correctly. We measure the expected increase in frontier (list of URIs to be crawled) size and wall-clock time required to archive resources, and investigate a way to recognize deferred representations to optimize crawler performance using a two-tiered approach that combines PhantomJS and Heritrix. Our efforts measure the crawling tradeoff between traditional archival tools and tools that can better archive JavaScript with headless browsing – a tradeoff that was anecdotally understood but not yet measured.

Throughout this paper we use Memento Framework terminology. Memento [37] is a framework that standardizes Web archive access and terminology. Original (or live web) resources are identified by URI-R, and archived versions of URI-Rs are called *mementos* and are identified by URI-M.

2. RELATED WORK

Archivability helps us understand what makes representations easier or harder to archive. Banos et al. created an algorithm to evaluate archival success based on adherence to standards for the purpose of assigning an archivability score [4]. In our previous work, we studied the factors influencing archivability, including accessibility standards and their impact on memento completeness, demonstrating that deviation from accessibility standards leads to reduced archivability [17]. We also demonstrated the correlation between the adoption of JavaScript and Ajax and the number of missing embedded resources in the archives [8].

Spaniol measured the quality of Web archives based on matching crawler strategies with resource change rates [10, 33, 34]. Ben Saad and Gançarski performed a similar study regarding the importance of changes on a page [5]. Gray and Martin created a framework for high quality mementos and assessed their quality by measuring the missing embedded resources [13]. In previous work, we measured the relative damage caused to mementos that were missing embedded resources to quantify the damage caused by missing resources loaded by JavaScript [7]. These works study quality, helping us understand what is missing from mementos.

David Rosenthal spoke about the difficulty of archiving representations enabled by JavaScript [25, 29]. Google has made efforts toward indexing deferred representations – a step in the direction of solving the archival challenges posed by deferred representations [6]. Google's indexing focuses on rendering an accurate representation for indexing and discovering new URIs, but does not completely solve the challenges to archiving caused by JavaScript. Archiving web resources and indexing representation content are different activities that have differing goals and processes.

Several efforts have studied client-side state. Mesbah et al. performed several experiments regarding crawling and indexing representations of Web pages that rely on JavaScript [19, 22]. These works have focused mainly on search engine indexing and automatic testing [20, 21] rather than archiving, but serve to illustrate the pervasive problem of deferred representations. Dincturk et al. constructed a model for crawling Rich Internet Applications (RIAs) by discovering all possible client-side states and identifying the simplest possible state machine to represent the states [11].

These prior works have focused on archival difficulties of crawling and indexing deferred representations, but have not explored the impact of archiving deferred representations on archival processes and crawlers. We measure the trade-off between speed and completeness of crawling techniques.

3. BACKGROUND

Web crawlers operate by starting with a finite set of seed URI-Rs in a frontier – or list of crawl targets – and add to the frontier by extracting URIs from the representations returned. Representations of Web resources are increasingly reliant on JavaScript and other client-side technologies to load embedded resources and control the activity on the client. Web browsers use a JavaScript engine to execute the client side code; Web crawlers traditionally do not have such an engine or the ability to execute client-side code because of the resulting loss of crawling speed. The client-side code can be used to request additional data or resources from servers (e.g., via Ajax) after the initial page load. Crawlers are unable to discover the resources requested via Ajax and, therefore, are not adding these URIs to their frontiers. The crawlers are missing embedded resources, which ultimately causes the mementos to be incomplete.

To mitigate the impact of JavaScript and Ajax on archivability, traditional crawlers that do not execute JavaScript (e.g., Heritrix) have constructed approaches for extracting links from embedded JavaScript to be added to crawl frontiers. Even though it does not execute JavaScript, Heritrix v. 3.1.4 does peek into the embedded JavaScript code to extract links where possible [16]. These processes rely on string matching and regular expressions to recognize URIs mentioned in the JavaScript. This is a sub-optimal approach because JavaScript may construct URIs from multiple strings during execution, leading to an incomplete URI extracted by the crawler.

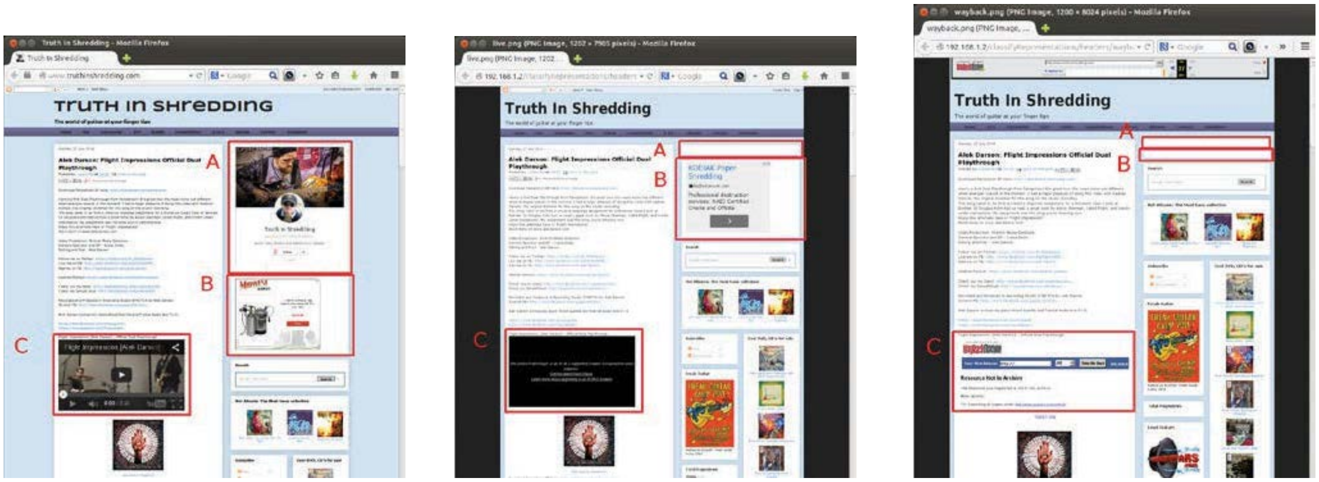
Because archival crawlers do not execute JavaScript, what is archived by automatic crawlers is increasingly different than what users experience. A solution to this challenge of archiving deferred representations is to provide crawlers with a JavaScript engine and allow headless browsing (i.e., allow a crawler to operate like a browser) using a technology such as PhantomJS. However, this change in crawling method impacts crawler performance, frontier size, and crawl time.

4. MOTIVATING EXAMPLES

To illustrate the challenge of archiving resources with deferred representations, we consider the resource at URI-R `http://www.truthinshredding.com/` and its mementos in Figure 1. We took a PNG snapshot of the live-Web resource as rendered in Mozilla Firefox (Figure 1(a)), the resource as loaded by PhantomJS (Figure 1(b)), and the memento created by Heritrix and viewed in a local installation of the Wayback Machine (Figure 1(c)). The title of the page "Truth in Shredding" appears in a different font in Figure 1(a) than in Figures 1(b) and 1(c) not due to a missing style sheet but rather an incompatibility of the font for the headless browser.

The live-Web resource loads embedded resources (annotated as A, B, and C) via JavaScript. Embedded Resource A is an HTML page loaded into an iframe. The original resource, $URI\text{-}R_A$, is

`https://apis.google.com/u/0/_/widget/render/page?usegapi=1&rel=publisher&href=%2F%2Fplus.google.com%2F110743665890542265089&width=430&hl=en-GB&origin=`



(a) The live resource at URI-R `http://www.truthinshredding.com/` loads A, B, and C via JavaScript.

(b) Using PhantomJS, the advertisement (B) and video (C) are found but the account frame (A) is missed.

(c) Using Heritrix, the embedded resources A, B, and C are missed.

Figure 1: Neither archival tool captures all embedded resources, but PhantomJS discovers the URI-Rs of two out of three embedded resources dependent upon JavaScript (B, C) while Heritrix misses all of them.

```
http%3A%2F%2Fwww.truthinshredding.com&gsrc=3p&ic
=1&jsh=m%3B%2F_%2Fscs%2Fapps-static...
```

The page loaded into the iframe uses JavaScript to pull the profile image into the page from $URI\text{-}R_{A_1}$

```
https://apis.google.com/_/scs/apps-static/_/ss/
k=oz.widget.-ynlzpp4csh.L.W.O/m=bdg/am=AAAAAJ
AwAA4/d=1/rs=AItRSTNrapszOr4y_tKMA1hZh6JM-g1haQ
```

Embedded Resource B is an advertisement that uses the JavaScript at $URI\text{-}R_{B_1}$

```
http://pagead2.googlesyndication.com/pagead/
show_ads.js
```

to pull in ads to the page. Embedded Resource C is a YouTube video that is embedded in the page using the following HTML for an iframe:

```
<iframe allowfullscreen="" frameborder="0" height=
"281" src="//www.youtube.com/embed/QyLl4Fd4cGA?rel
=0" width="500"></iframe>.
```

PhantomJS does not load Embedded Resource A, potentially because the host resource completes loading before the page embedded in the iframe can finish loading. PhantomJS stops recording embedded URIs and monitoring the representation after a page has completed loading, and Embedded Resource A executes its JavaScript to load the profile picture after the main representation has completed the page load[1]. PhantomJS does discover the advertisement (Embedded Resource B) and the YouTube video (Embedded Resource C). Even though the headless browser used by PhantomJS does not have the plugin necessary to display the video, the URI-R is still discovered by PhantomJS.

Heritrix fails to identify the URI-Rs for the Embedded Resources A, B, and C. When the memento created by Heritrix is loaded by the Wayback Machine, Embedded Resources A, B, and C are missing. This is attributed to Heritrix, which does not discover the URI-Rs for these resources during the crawl. When viewing the memento through the Wayback Machine, the JavaScript responsible for loading the embedded resources is executed resulting in either a zombie resource (*prima facie violative*) or HTTP 404 response (incomplete) for the embedded URI.

Heritrix's inability to discover the embedded URI-Rs could be mitigated by utilizing PhantomJS during the crawl. However, this raises many questions, most notably: How much slower will the crawl time be? How many additional embedded resources could it recover and potentially need to store? Can we optimize the crawl approach based on the detection of deferred representations? Our investigation into these questions will assess the feasibility of combining Heritrix with PhantomJS to balance the speed of Heritrix with the completeness of PhantomJS.

5. COMPARING CRAWLS

We designed an experiment to measure the performance differences between a command-line archival tool (wget [12]), a traditional crawler (the Internet Archive's Heritrix Crawler [23, 30]), and a headless browser client (PhantomJS). Neither Heritrix nor wget execute the client-side JavaScript, while PhantomJS *does* execute client-side JavaScript.

We constructed a 10,000 URI-R dataset by randomly generating a Bitly URI and extracting its redirection target

[1] PhantomJS scripts can be written to avoid this race-condition using longer timeouts or client-side event detection, but this is outside the scope of this paper.

(identical to the process used to create the Bitly data subset in [1]). We split the 10,000 URI dataset into 20 sets of 500 seed URI-Rs and used wget, Heritrix, and PhantomJS to crawl each set of seed URI-Rs. We repeated each crawl ten times to establish an average performance, resulting in ten different crawls of the 10,000 URI dataset (executing the crawl one of the 500-URI sets at a time) with wget, Heritrix, and PhantomJS. We measured the increase in frontier size ($|F|$) and the URIs per second (t_{URI}) to crawl the resource.

While Heritrix provides a user interface that identifies the crawl frontier size, PhantomJS and wget do not. We calculate the frontier size of PhantomJS by counting the number of embedded resources that PhantomJS requests when rendering the representation. We calculate the frontier size of wget by executing a command[2] that records the HTTP GET requests issued by wget during the process of mirroring a web resource and its embedded resources. We consider the frontier size to be the total number of resources and embedded resources that wget attempts to download.

We began a crawl of the same 500 URI-Rs using wget, Heritrix, and PhantomJS simultaneously to mitigate the impact of live Web resources changing state during the crawls. For example, if the representation changes (such as includes new embedded resources) in between the times wget, PhantomJS, and Heritrix perform their crawls, the number or representations of embedded resources may change and therefore the representation influenced the crawl performance, not the crawler itself.

We crawled live-Web resources because mementos inherit the limitations of the crawler used to create them. Depending on crawl policies, a memento may be incomplete and different than the live resource. The robots.txt protocol [27, 35], breadth- versus depth-first crawling, or the inability to crawl certain representations (like deferred representations as we discuss in this paper) can all influence the mementos created during a crawl.

5.1 Crawl Time by URI

To better understand how crawl times of wget, PhantomJS, and Heritrix differ, we determined the time needed to execute a crawl. Heritrix has a browser-based user interface that provides the URIs/second (t_{URI}) metric. We collected this metric from the Web interface for each crawl. We used Unix system times to calculate the crawl time for each PhantomJS and wget crawl by determining the start and stop times for dereferencing each resource and its embedded resources. We compare the wget, PhantomJS, and Heritrix crawl times per URI in Figure 2 and Table 1. Heritrix outperforms PhantomJS, crawling 2.065 URIs/s while PhantomJS crawls 0.170 URIs/s and wget crawls 0.864 URIs/s. Heritrix crawls, on average, 12.13 times faster than PhantomJS and 2.39 times faster than wget.

The performance difference comes from two aspects of the crawl. First, Heritrix executes crawls in parallel with multiple threads being managed by the Heritrix software – this is not possible with PhantomJS on a single core machine since PhantomJS requires access to a headless browser and its associated JavaScript engine, and parallelization will result in process and threading conflicts. Second, Heritrix does not execute the client-side JavaScript and only adds URIs that are extracted from the Document Object Model (DOM), embedded style sheets, and other resources to its frontier.

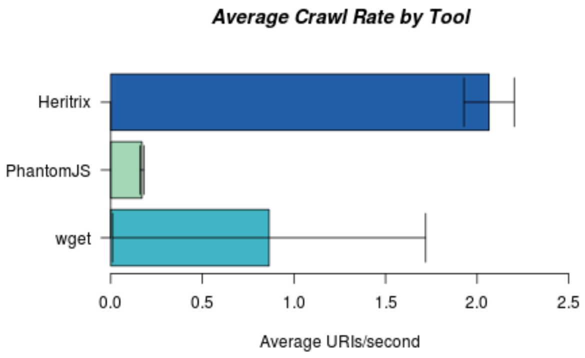


Figure 2: Heritrix crawls 12.13 times faster than PhantomJS. The error lines indicate the standard deviation across all ten runs.

5.2 URI Discovery and Frontier Size

We performed a string-matching de-duplication (that is, removing duplicate URIs) to determine the true frontier size ($|F|$).

Crawler	Crawl time		Frontier Size	
	$\overline{t_{URI}}$	$s_{t_{URI}}$	$\overline{\lvert F \rvert}$	$s_{\lvert F \rvert}$
wget	0.864	0.855	129,443	3,213.65
Heritrix	2.065	0.137	302,961	1,219.82
PhantomJS	0.170	0.001	531,484	2,036.92

Table 1: Mean and standard deviation of crawl time (URIs/s) and frontier size for wget, Heritrix, and PhantomJS crawls of 10,000 seed URIs.

As shown in Figure 3 and in Table 1, we found that PhantomJS discovered and added 1.75 times more URI-Rs to its frontier than Heritrix, and 4.11 times more URI-Rs than wget. Per URI-R, PhantomJS loads 19.7 more embedded resources than Heritrix and 32.4 more embedded resources than wget. The superior PhantomJS frontier size is attributed to its ability to execute JavaScript and discover URIs constructed and requested by the client-side scripts.

However, raw frontier size is not the only performance metric for assessing the quality of the frontier. PhantomJS and Heritrix discover some of the same URIs, while PhantomJS discovers URIs that Heritrix does not and Heritrix discovers URIs that PhantomJS does not. We measured the union and intersection of the Heritrix and PhantomJS frontiers. As shown in Figure 4(a), per 10,000 URI-R crawl Heritrix finds 39,830 URI-Rs missed by PhantomJS on average, while PhantomJS finds 194,818 URI-Rs missed by Heritrix per crawl on average. PhantomJS and Heritrix find 63,550 URI-Rs in common between the two crawlers. The wget crawl

[2] We executed `wget -T 40 -o outfile -p -O headerFile [URI-R]` which downloads the target URI-R and all embedded resources and dumps the HTTP traffic to `headerFile`.

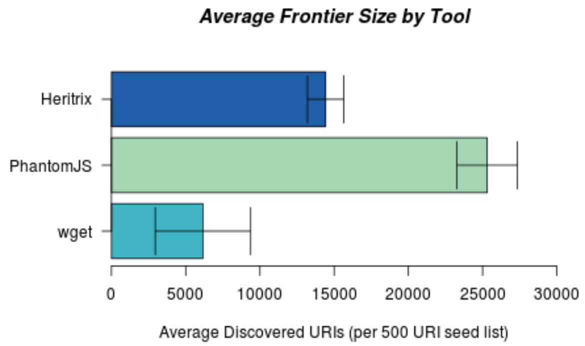
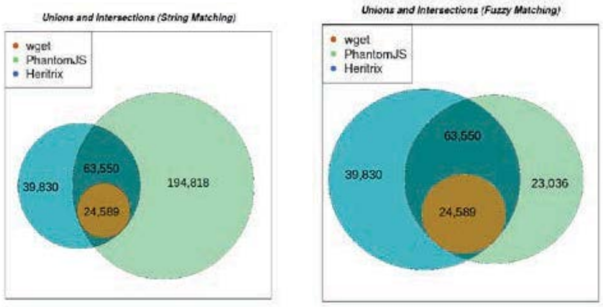


Figure 3: PhantomJS discovers 1.75 times more embedded resources than Heritrix and 4.11 times more resources than wget. The averages and error lines indicate the standard deviation across all ten runs.



(a) A portion of Heritrix, PhantomJS, and wget frontiers overlap. PhantomJS and Heritrix identify URIs that the others do not.

(b) The frontier of URI-Rs unique to PhantomJS shrinks when only considering the host and path aspects (Base Policy for matching) of the URI-R.

Figure 4: Heritrix, PhantomJS, and wget frontiers as an Euler Diagram. The overlap changes depending on how duplicate URIs are identified.

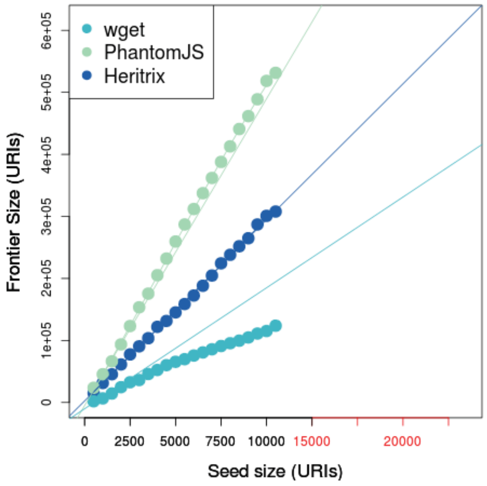


Figure 5: Frontier size grows linearly with seed size.

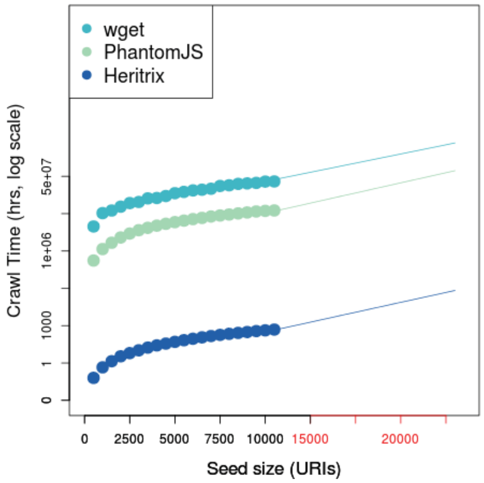


Figure 6: Crawl speed is dependent upon frontier size.

resulted in a frontier of 24,589 URI-Rs, which was a proper subset of both the Heritrix and PhantomJS frontiers.

This analysis shows that PhantomJS finds 19.70 more embedded resources per URI than Heritrix (Figure 5). Heritrix runs 12.13 times faster than PhantomJS (Figure 6). Note that the red axis in Figures 5 and 6 are unmeasured and only projections of the measured trends, with the projections predicting the performance as the seed list size grows.

5.3 Frontier Properties

During the PhantomJS crawls, we observed that PhantomJS discovers session-specific URI-Rs that Heritrix misses and Heritrix discovers Top Level Domains (TLDs) that PhantomJS misses, presumably from Heritrix's inspection of JavaScript. For example:

```
http://dg.specificclick.net/?y=3&t=h&u=http%3A%2F%2
Fmisscellania.blogspot.com%2Fstorage%2F
Twitter-2.png...
```

from PhantomJS versus

```
http://dg.specificclick.net/
```

from Heritrix. The uniquely Heritrix URI-Rs are potentially the base of a URI to be further built by JavaScript. Because PhantomJS only discovers URIs for which the client issues HTTP requests, this URI-R is not discovered by PhantomJS. To determine the nature of the differences between the Heritrix and PhantomJS frontiers, we analyzed the union and intersection between the URI-Rs in the frontiers using different matching policies (Figure 4(b)).

During a crawl of 500 URI-Rs by PhantomJS, 19,022 URI-Rs were added to the frontier for a total of 19,522 URI-Rs in the frontier. We also captured the content body (the returned entity received when dereferencing a URI-R from the frontier) and recorded its MD5 hash value. We used the hash value to identify duplicate representations during the crawl. To determine duplication between URIs, we used five matching policies to determine the duplication within the frontier (Table 2). In other words, we identify cases in which the URIs are different but the content is the same, similar to the methods used by Sigurðsson [31, 32].

The *No Trim* policy uses strict string matching of the URI-Rs to detect duplicates. The *Base Trim* policy trims all parameters from the URI. For example, the URI

```
http://example.com/folder/index.html?param=value
```

would be trimmed to

```
http://example.com/folder/index.html
```

The *Origin Trim* policy eliminates all parameters and associated values that reference a referring source, such as `origin`, `callback`, `domain`, or `referrer`. These parameters are often associated with a value including the top level domain of the referring page. Frequent implementers include Google Analytics or ad services.

The *Session Trim* policy eliminates all parameters and their associated values that reference a session. For example, the parameters such as `session`, `sessionid`, `token_id`, etc. are all removed from the URI-R before matching. These parameters are often used by ad services or streaming media services to identify browsing sessions for tracking and revenue generation purposes.

The *HTTP Trim* policy removes all parameters with values that mention a URI. Ad services, JavaScript files, and other statistics tracking services frequently utilize these parameters. For example, the URI

```
http://example.com/folder/index.html?param=value
&httpParam=http://www.test.com/
```

would be trimmed to

```
http://example.com/folder/index.html?param=value
```

We used the five trimming policies to detect duplicates in the frontiers constructed by PhantomJS in one of the crawls of 500 URI-Rs. At the end of the crawl, PhantomJS had a frontier of 19,522 URI-Rs. Using the MD5 hash of the representations, we determined that this set had 8,859 duplicate representations. With the trimmed URI and the MD5 hash of the entity, we can compare the identifiers and the returned entities for duplication.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Number of Classifications}} \qquad (1)$$

Trim Type	URI Duplicates	URI and Entity Duplicates	Accuracy
No Trim	6,469	4,684	0.68
Other Trim	6,933	2,810	0.62
Origin Trim	7,078	4,749	0.68
Base Trim	10,359	5,191	0.56
Session Trim	8,159	4,921	0.64
HTTP Trim	7,315	4,868	0.67

Table 2: Detected duplicate URIs, entity bodies, and the overlap between the two using the five URI string trimming policies.

$$\text{F-Measure} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (2)$$

For each of the 19,522 URIs in the frontier and their associated entity hash values, we determined the trimmed URI string and the duplications of URIs in the frontier and the number of duplicate URIs that also had a duplicate entity body (Table 2). We calculated the accuracy (Equation 1)[3] of each trim policy using the number of URIs with the same entity hash and URI as a true positive (TP), the number of URIs that had neither a duplicate URI nor a duplicate entity body as a true negative (TN), and the set of all positives and negatives (P + N) as the total number of URIs (19,522).

The *Base Trim* and *No Trim* policies had identical accuracy ratings (0.68). The *Base Trim* policy identified the most URI duplicates, and is used to determine the overlap between the Heritrix and PhantomJS frontiers.

Using the *Base Trim* policy to only consider the host and path (e.g., `http://pubads.g.doubleclick.net/gampad/ads`) of the PhantomJS and Heritrix frontiers, PhantomJS identifies 376,578 URI-Rs added to the frontier, 199,761 (55%) of which are duplicates of the discovered URIs. If we consider only the host and path of the PhantomJS URIs, the Euler Diagram of PhantomJS and Heritrix frontiers is more evenly matched (Figure 4(b)).

5.4 Deferred vs. Non-Deferred Crawls

To isolate the impact of resources with deferred representations on crawl performance, we manually classified 200 URI-Rs from our set of 10,000 URI-Rs as having deferred representations and another 200 as having non-deferred representations. We crawled each of the deferred and non-deferred sets of URI-Rs with PhantomJS and Heritrix.

During the crawl of the non-deferred set, PhantomJS crawled t_{URI}=0.255 URIs/s while Heritrix crawled t_{URI}=1.34 URIs/s, 5.25 times faster than PhantomJS. Heritrix uncovered 1,044 URI-Rs to add to the frontier, while PhantomJS discovered 403 URI-Rs to add to the frontier. This phenomenon of

[3] Accuracy is defined as the number of correctly classified instances divided by the test set size (Equation 1). F-Measure extends accuracy to consider the harmonic mean of precision and recall (Equation 2).

Heritrix having a larger frontier than PhantomJS is due to Heritrix's policy of looking into the JavaScript files to extract URIs found in the code – the URI-Rs discovered by Heritrix are top-level domains listed in the JavaScript that may be used to construct URIs at run time (e.g., appending a username or timestamp to the URI) or not used by JavaScript at all (e.g., a URI that exists in un-executed code).

During the crawl of the deferred set, PhantomJS crawled t_{URI}=0.5. Heritrix ran t_{URI}=12.56, 25.12 times faster than PhantomJS. Heritrix added 3,206 URIs to the frontier, while PhantomJS added 3,436 URIs to the frontier. PhantomJS adds more URIs to the frontier despite Heritrix's introspection on the JavaScript of each crawl target. This result is due to PhantomJS's execution of JavaScript on the client.

We observe that the PhantomJS frontier outperforms the Heritrix frontier during the deferred crawl. Heritrix crawls URIs faster than PhantomJS on each of the deferred and non-deferred crawls, but far exceeds the speed of PhantomJS during the deferred crawl.

6. CLASSIFYING REPRESENTATIONS

In practice, archival crawlers such as Heritrix would be able to identify URI-Rs that have low archivability in real-time. Heritrix currently does not have such an automatic capability. Archive-It, for example, uses a manually curated list of URIs that have deferred representatiosn and uses Umbra [26] to crawl them.

The ability to determine the archivability of a resource will allow Heritrix to assign the URI-R to either the faster, traditional Heritrix crawler or the slower, PhantomJS (or other JavaScript-enabled crawler). By enabling this two-tiered approach to crawling, the archival crawlers can achieve maximum performance by utilizing the heavy-duty JavaScript-capable crawlers for only those that need it. However, this approach requires the ability to, in real-time, recognize or predict a deferred representation.

Even though our goal is to detect whether or not representations are dependent on JavaScript, the simple presence of JavaScript is not a sufficient indicator of a deferred representation. In our set of URI-Rs, the resources with deferred representations had, on average, 21.98 embedded script tags or files, while the resources with non-deferred representations had 5.3 script tags or files. Of those resources with deferred representations, 84.1% had at least one script tag, while 49.5% of the non-deferred representations had at least one script tag. Because of the ubiquity of JavaScript in both deferred and non-deferred representations, we opted for a more complex feature vector to represent the features of the representations.

In an effort to predict whether or not a representation would be deferred, we constructed a feature vector of DOM attributes and features of the embedded resources. We used Weka [14] to classify the resources on subsets of the feature vectors to gauge their performance. We extracted the following feature vector:

1. **Ads**: Using a list of known advertisement domains, we determined whether or not a representation would load an ad based on DOM and JavaScript analysis.
2. **Script Tags**: We counted the number of script tags with JavaScript, both in files and embedded code.
3. **Interactive Elements**: We counted the number of DOM elements that have JavaScript events attached to them (e.g., `onclick`, `onload`).
4. **Ajax (in JavaScript)**: To estimate the number of Ajax calls (e.g., `$.get()`, `XmlHttpRequest`) we counted the number of occurrences of Ajax requests in the embedded external and independent JavaScript files.
5. **Ajax (in HTML)**: To estimate the number of Ajax calls (e.g., `$.get()`, `XmlHttpRequest`) we counted the number of occurrences of Ajax requests in Script tags embedded in the DOM.
6. **DOM Modifications**: We counted the number of times JavaScript made a modification of the DOM (e.g., via the `appendChild()` function) to account for DOM modifications after the initial page load.
7. **JavaScript Navigation**: We counted the occurrences of JavaScript redirection and other navigation functions (e.g., `window.location` calls).
8. **JavaScript Storage**: We count the number of JavaScript references to storage elements on the client (e.g., cookies) as an indication of client-controlled state.
9. **Found, Same Domain**: Using PhantomJS, we counted the number of embedded resources originating from the URI-R's top level domain (TLD) that were successfully dereferenced (i.e., returned an HTTP 200).
10. **Missed, Same Domain**: Using PhantomJS, we counted the number of embedded resources originating from the URI-R's TLD that were not successfully dereferenced (i.e., returned a class HTTP 400 or 500).
11. **Found, Different Domain**: Using PhantomJS, we counted the number of embedded resources originating outside of the URI-R's TLD that were successfully dereferenced (i.e., returned an HTTP 200).
12. **Missed, Different Domain**: Using PhantomJS, we counted the number of embedded resources originating outside of the URI-R's TLD that were unsuccessfully dereferenced (i.e., a class 400 or 500 HTTP response).

We manually sampled 440 URI-Rs (from our collection of 10,000, including the same 400 from Section 5.4) and classified the representations as deferred or non-deferred, with 200 training and 20 test URI-Rs for each based on whether or not their representations were dependent upon JavaScript.

Using PhantomJS, we collected the 12 features required for a feature vector for each of our 440 URI-Rs. Using Weka, we ran each classifier on the feature vectors. Rotation Forests [28] performed the best of any of the standard Weka classifiers for any of our datasets.

We used three subsets of the feature vector to investigate the best method of predicting deferred representations. We selected attributes 1-8 to represent DOM features. We selected attributes 9-12 as embedded resource attributes (the attributes we extract if we load and monitor the embedded resources). Together, attributes 1-12 make up the entire dataset. We use the feature sets to train and test our classifier via 10-fold cross validation. We use the same three data

Actual Classification	**Predicted Classification**	
	Deferred	Non-Deferred
Deferred	182	38
Non-Deferred	58	166

Table 3: Confusion matrix for the entire feature vector (F-Measure = 0.791).

Actual Classification	**Predicted Classification**	
	Deferred	Non-Deferred
Deferred	179	41
Non-Deferred	47	173

Table 4: Confusion matrix for the resource features (features 9-12 of the vector; F-Measure = 0.844).

subsets and provide a confusion matrix of each set including the entire feature vector (Table 3), resource feature vector (Table 4), and DOM feature vector (Table 5).

The accompanying statistics for the classifications are shown in Table 6. With only the DOM features, the test set is accurately classified representations as deferred or non-deferred 79% of the time. If we combine the DOM and resource feature sets to create the full feature set, we can correctly classify representations 81% of the time.

After a URI is dereferenced and a representation is returned, we can determine whether or not the representation is deferred with 79% accuracy. If we also dereference the URIs for the embedded resources and monitor the HTTP status codes, we can increase, albeit minimally, the accuracy of the prediction to 81% of the time. However, crawling with PhantomJS is much more expensive when executed properly. Due to this minimal improvement and much higher cost to measure, the feature extraction will be limited to the DOM classification. With a negligible impact on performance, our classifier is able to identify deferred representations using the DOM crawled by Heritrix with 79% accuracy.

7. TWO-TIERED CRAWLING

To benefit from the increased crawl frontier size of PhantomJS while maintaining the performance of Heritrix, we propose a tiered crawling approach in which PhantomJS is used to crawl only resources with deferred representations. A tiered approach to crawling would allow an archive to simultaneously benefit from the frontier size of PhantomJS and the speed of Heritrix. Table 7 provides a summary of the extrapolated crawl speed and discovered frontier size of each crawler. While the test environment used a single system, a production environment should expect to see performance improvements with additional resources. PhantomJS crawls are not run in parallel, and additional nodes for PhantomJS threads will further improve performance.

Actual Classification	**Predicted Classification**	
	Deferred	Non-Deferred
Deferred	168	52
Non-Deferred	41	179

Table 5: Confusion matrix for the DOM features (features 1-8 of the vector; F-Measure = 0.806).

Crawl Strategy	Crawl Time (hrs)	Crawl Rate (t_{URI})	Frontier Size ($\|F\|$)
wget	416.16	0.864	129,443
Heritrix	407.53	2.065	302,961
PhantomJS	8,684.38	0.170	531,484
Heritrix + PhantomJS	9,100.54	0.152	537,609
Heritrix + PhantomJS with Classifier	6,495.23	0.196	458,815

Table 7: A summary of *extrapolated* performance (based on our calculations) of single- and two-tiered crawling approaches.

We have described the operation of crawls with wget, Heritrix, and PhantomJS in Sections 5.1 and 5.4 with wget serving as a baseline to which Heritrix and PhantomJS can be compared but wget is not part of the archival workflow we investigate. To reiterate, Heritrix crawls much more quickly than PhantomJS, while PhantomJS discovers many more embedded resources required to properly construct a representation. Optimally during a crawl, Heritrix would dereference a URI-R and run the resulting DOM through the classifier to determine whether or not the representation will be deferred (with 79% accuracy, as discussed in Section 6). If the representation is predicted to be deferred, PhantomJS should also be used to crawl the URI-R and add the newly discovered URI-Rs to the Heritrix frontier.

Heritrix should be used to crawl all URI-Rs in the frontier because the DOM is required to classify a representation as deferred. Since Heritrix is the fastest crawler, it should be used to dereference the URI-Rs in the frontier and retrieve the DOM of the resource for classification. Subsequently, only if the representation is classified as deferred will PhantomJS be used to crawl the resource to ensure the maximum amount of embedded resources are retrieved.

In a naive two-tiered crawl strategy that will discover the most embedded URI-Rs and create the largest frontier, Heritrix and PhantomJS should both crawl each URI-R regardless of whether the representation can be classified as deferred or non-deferred. This creates a crawl that is expected to be 13.5 times slower than simply using Heritrix, but is expected to discover 1.77 times more URI-Rs than using only Heritrix. This would ensure that 100% of all resources with deferred representations would be crawled with both Heritrix and PhantomJS. However, we want to limit the use of PhantomJS to minimize the performance impacts it has on the crawl speed.

If we include the classifier to predict when PhantomJS should be used or when Heritrix will be a suitable tool, the two-tiered approach is expected to run 10.5 times slower and is expected to discover 1.5 times more URI-Rs than only Heritrix. This crawl policy balances the trade-offs between speed and larger frontier size by using the classifier to indicate when to use PhantomJS to crawl resources with deferred representations.

Features	Classification	Accuracy	F-measure	Precision	Recall
DOM Features Only	Deferred Non-deferred	79%	79%	78% 76%	81% 80%
DOM & Resource Features	Deferred Non-deferred	81%	82%	79% 90%	81% 80%

Table 6: Classification success statistics for DOM-only and DOM and Resource feature sets.

To validate this expected calculation, we classified our 10,000 URI-R dataset, which produced 5,187 URI-Rs classified as having deferred representations, and 4,813 as having non-deferred representations. We used PhantomJS to crawl the URI-Rs classified as deferred, and only Heritrix to crawl the URI-Rs classified as non-deferred. The results of the crawls are detailed in Table 8.

Crawler	URI-R Set	Seed Size	Frontier Size	Crawl Time (hrs)
P	Deferred	5,187	311,903	84.9
H	Non-deferred	4,813	124,728	23.6
H	Deferred	5,187	171,499	26.7
P	All URI-Rs	10,000	438,388	686
H	All URI-Rs	10,000	275,234	48.3
Two-tier	All URI-Rs	10,000	399,202	133

Table 8: A simulated two-tiered crawl showing that the frontier sizes can be optimized while mitigating the performance impact of PhantomJS's (P) crawl speed vs Heritrix's (H).

In this table, we show that PhantomJS creates a frontier of 438,388, 1.6 times larger than that of Heritrix. However, PhantomJS crawls 14 times slower than Heritrix. If we perform a tiered crawl in which PhantomJS is responsible for crawling only deferred representations, we can crawl 5.2 times faster than using only PhantomJS (but 2.7 times slower than the Heritrix-only approach) while creating a frontier 1.8 times larger than using only Heritrix. As a result, we can maximize the frontier size, mitigate the impacts of JavaScript on crawling, and mitigate the impact of the reduced crawl speeds when using a tiered crawling approach.

8. CONCLUSIONS

In this paper, we measured the differences in crawl speed and frontier size of wget, PhantomJS, and Heritrix. While PhantomJS was the slowest crawler, it provided the largest crawl frontier due to its ability execute client-side JavaScript to discover URIs missed by Heritrix and wget. Heritrix was the fastest crawler. We also proposed a tiered approach to crawling in which a classifier determines whether to crawl a resource with PhantomJS to reap the URI discovery benefits of the specialized crawler where appropriate.

This work lays the foundation for a two-tiered crawling approach and helps predict the performance of future archival workflows. We know that PhantomJS finds 19.70 more embedded resources per URI and Heritrix runs 12.13 times faster than PhantomJS, meaning the crawler should avoid crawling URI-Rs with non-deferred representations to maintain an optimal performance trade-off. We understand that PhantomJS is required to discover the embedded resources needed to complete a deferred representation that Heritrix cannot discover. This has a performance detriment to run time, but offers a benefit of more complete mementos and a larger frontier for crawling. We also found that 53% of URIs discovered by PhantomJS are duplicates if we remove session-specific URI parameters.

Using DOM features we can accurately predict deferred and non-deferred representations 79% of the time. Using this classification, deferred representations can be crawled by PhantomJS to ensure all embedded resources are added to the crawl frontier.

If using a multi-tiered approach to crawling, archives can leverage the benefits of PhantomJS and Heritrix simultaneously. That is, using a deferred representation classifier, archives can use PhantomJS for deferred representations and Heritrix for non-deferred representations. Using a tiered crawling approach, we showed that crawls will run 5.2 times faster than using only PhantomJS, create a frontier 1.8 times larger than using only Heritrix. This crawl strategy mitigates the impact of JavaScript on archiving while also mitigating the reduced crawl speed of PhantomJS.

Our future work will include a framework for archiving deferred representations, along with a measurement of the archival improvement when implementing a deferred representation crawler.

9. ACKNOWLEDGMENTS

This work supported in part by the NEH (HK-50181).

References

[1] S. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How much of the Web is archived? In *JCDL 2011*, pages 133–136, 2011.

[2] S. Ainsworth, M. L. Nelson, and H. Van de Sompel. A Framework for Evaluation of Composite Memento Temporal Coherence. Technical Report arXiv:1402.0928, 2014

[3] Y. Alnoamany, A. Alsum, M. Weigle, and M. Nelson. Who and What Links to the Internet Archive. In *TPDL 2013*, pages 346–357. 2013.

[4] V. Banos, K. Yunhyong, S. Ross, and Y. Manolopoulos. CLEAR: a Credible Method to Evaluate Website Archivability. In *iPRES 2013*, 2013.

[5] M. Ben Saad and S. Gançarski. Archiving the web using page changes patterns: A case study. In *JCDL 2011*, pages 113–122, 2011.

[6] J. F. Brunelle. Google and JavaScript. http://ws-dl.blogspot.com/2014/06/2014-06-18-google-and-javascript.html, 2014.

[7] J. F. Brunelle, M. Kelly, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. Not All Mementos Are Created Equal: Measuring The Impact Of Missing Resources. In *JCDL 2014*, pages 321 – 330, 2014.

[8] J. F. Brunelle, M. Kelly, M. C. Weigle, and M. L. Nelson. The Impact of JavaScript on Archivability. *International Journal on Digital Libraries*, pages 1–23, 2015.

[9] J. F. Brunelle and M. L. Nelson. Zombies in the archives. http://ws-dl.blogspot.com/2012/10/2012-10-10-zombies-in-archives.html, 2012.

[10] D. Denev, A. Mazeika, M. Spaniol, and G. Weikum. SHARC: framework for quality-conscious web archiving. *Proceedings of the 35th International Conference on VLDB*, 2:586–597, August 2009.

[11] M. E. Dincturk, G.-V. Jourdan, G. V. Bochmann, and I. V. Onut. A Model-Based Approach for Crawling Rich Internet Applications. *ACM Transactions on the Web*, 8(3):19:1–19:39, July 2014.

[12] GNU. Introduction to GNU Wget. http://www.gnu.org/software/wget/, 2013.

[13] G. Gray and S. Martin. Choosing a sustainable web archiving method: A comparison of capture quality. *D-Lib Magazine*, 19(5), May 2013.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11(1):10–18, 2009.

[15] B. A. Howell. Proving Web History: How to Use the Internet Archive. *Journal of Internet Law*, 9(8):3–9, February 2006.

[16] P. Jack. ExtractorHTML Extract-JavaScript. https://webarchive.jira.com/wiki/display/Heritrix/ExtractorHTML+extract-javascript, 2014.

[17] M. Kelly, J. F. Brunelle, M. C. Weigle, and M. L. Nelson. On the Change in Archivability of Websites Over Time. In *TPDL 2013*, pages 35–47, 2013.

[18] C. C. Marshall and F. M. Shipman. On the Institutional Archiving of Social Media. In *JCDL 2012*, pages 1–10, 2012.

[19] A. Mesbah, E. Bozdag, and A. van Deursen. Crawling Ajax by inferring user interface state changes. In *ICWE 2008*, pages 122 –134, 2008.

[20] A. Mesbah and A. van Deursen. Migrating multi-page web applications to single-page Ajax interfaces. In *CSMR 2007*, pages 181–190. 2007.

[21] A. Mesbah and A. van Deursen. Invariant-based automatic testing of Ajax user interfaces. In *ICSE 2009*, pages 210–220. 2009.

[22] A. Mesbah, A. van Deursen, and S. Lenselink. Crawling Ajax-Based Web Applications Through Dynamic Analysis of User Interface State Changes. *ACM Transactions on the Web*, 6(1):3:1–3:30, March 2012.

[23] G. Mohr, M. Kimpton, M. Stack, and I. Ranitovic. Introduction to Heritrix, an archival quality web crawler. In *Proceedings of the 4th IWAW*, September 2004.

[24] K. C. Negulescu. Web Archiving @ the Internet Archive. Presentation at the 2010 Digital Preservation Partners Meeting, 2010 http://1.usa.gov/1GkRUDE.

[25] NetPreserver.org. IIPC Future of the Web Workshop – Introduction & Overview. http://netpreserve.org/sites/default/files/resources/OverviewFutureWebWorkshop.pdf, 2012.

[26] S. Reed. Introduction to Umbra. https://webarchive.jira.com/wiki/display/ARIH/Introduction+to+Umbra, 2014.

[27] Robots.txt. The Web Robots Page. http://www.robotstxt.org/, 2014.

[28] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso. Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1619–1630, Oct. 2006.

[29] D. S. H. Rosenthal. Talk on Harvesting the Future Web at IIPC2013. http://blog.dshr.org/2013/04/talk-on-harvesting-future-web-at.html, 2013.

[30] K. Sigurðsson. Incremental crawling with Heritrix. In *Proceedings of the 5th IWAW*, September 2005.

[31] K. Sigurðsson. The results of URI-agnostic deduplication on a domain crawl. http://kris-sigur.blogspot.com/2014/12/the-results-of-uri-agnostic.html, 2014.

[32] K. Sigurðsson. URI agnostic deduplication on content discovered at crawl time. http://kris-sigur.blogspot.com/2014/12/uri-agnostic-deduplication-on-content.html, 2014.

[33] M. Spaniol, D. Denev, A. Mazeika, G. Weikum, and P. Senellart. Data quality in web archiving. In *Proceedings of the 3rd Workshop on Information Credibility on the Web*, pages 19–26. 2009.

[34] M. Spaniol, A. Mazeika, D. Denev, and G. Weikum. Catch me if you can: Visual Analysis of Coherence Defects in Web Archiving. In *Proceedings of The 9th IWAW*, pages 27–37, 2009.

[35] Y. Sun, Z. Zhuang, and C. L. Giles. A large-scale study of robots.txt. In *WWW 2007*, pages 1123–1124, 2007.

[36] B. Tofel. 'Wayback' for Accessing Web Archives. In *Proceedings of the 7th IWAW*, 2007.

[37] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth, and H. Shankar. Memento: Time Travel for the Web. Technical Report arXiv:0911.1112, 2009.

Experiment, Document & Decide: a Collaborative Approach to Preservation Planning at the BnF

Bertrand Caron
Department of Bibliographic and Digital Information
Bibliothèque nationale de France
bertrand.caron@bnf.fr

Thomas Ledoux
Department of Information Technology
Bibliothèque nationale de France
thomas.ledoux@bnf.fr

Stéphane Reecht
Department of Conservation and Preservation
Bibliothèque nationale de France
stephane.reecht@bnf.fr

Jean-Philippe Tramoni
Department of Information Technology
Bibliothèque nationale de France
jean-philippe.tramoni@bnf.fr

ABSTRACT

The National Library of France (BnF) has recently implemented a new module for its Scalable Preservation and Archiving Repository (SPAR) to set up preservation strategies based on formats, agents, workflows, tools and tests, and managed as reference packages in the Archive.

This module aims to fulfill an objective: for SPAR to be fully self-documented. Formats, agents and workflows are formally described and preserved along with the Information packages in which such elements are involved. Although this was a feature that was included from the beginnings of SPAR, the new Preservation Planning module aims to provide a tool that can more easily build these reference packages and that will more closely involve domain experts and the IT department in the processes of preservation planning. But the main innovation lies in the documentation of decisions that directed their selection as standards in SPAR: test data are now preserved as a new kind of reference package.

General Terms

Preservation strategies and workflows; innovative practice.

Keywords

Preservation planning, decision documentation, community involvement.

1. INTRODUCTION

Since the operational launch of SPAR in 2010, the BnF has had to face a growing diversity of digital documents (heritage digitization in 2010, third-party archiving in 2011, web archiving in 2013, legal deposit of ebooks planned for 2015). Ingest and preservation of these specific materials led the BnF to implement many different workflows involving characterization, processing and transformation tools.

The BnF felt the urge to record the choices made about these operations not only within the system logs but also within the repository itself. From data objects on which tests were performed to results of said tests using a software tool, every step explaining the decisions that led experts to carry out a specific preservation plan has to be preserved.

Following the path initiated by the experimental tool Plato [1] and based on discussions with various communities, the "Preservation planning" module was developed to address this specific need. Although all activities of the Preservation planning OAIS entity were not taken over, its first version allows experts to develop preservation strategies and standards and keep track of their elaboration.

The module is provided with a user-friendly interface and several levels of authorization; its objective is to foster collaborative work between experts from different departments of the library.

2. WHY A PRESERVATION PLANNING MODULE?

2.1 What Does OAIS Say?

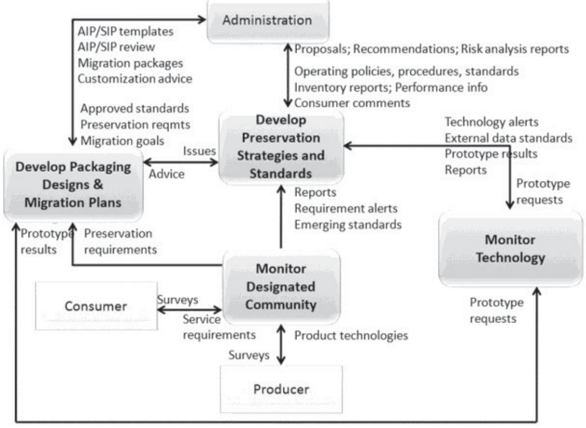


Figure 1. Functions of the 'Preservation planning' entity (source: Reference model for an Open Archival Information System (OAIS) [2])

As defined by the OAIS standard, the Preservation Planning Functional Entity "provides the services and functions for monitoring the environment of the OAIS, providing recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, the Designated Community over the Long

Term, even if the original computing environment becomes obsolete.”

Monitoring the environment of the archive is achieved through two functions: Monitor Designated Community and Monitor Technology. The former calls for interactions with members of the community in order to track changes in their service requirements or product technologies. The latter requires performing surveillance on emerging standards or technologies. Any changes are reported to the two other functions of this entity which are responsible for defining, developing and validating preservation plans and appropriate tools (see [Figure 1]).

The Develop Preservation Strategies and Standards function “is responsible for developing and recommending strategies and standards, and for assessing risks, to enable the Archive to make informed tradeoffs as it establishes standards, sets policies, and manages its system infrastructure.” In response to the reports about identified changes in the environment of the archive, this function will have to estimate possible updates in archive operations, including policies, procedures, standards and tools. This evaluation may require prototyping and testing of these updates such as: SIP/AIP templates, submission requirements, new or modified file formats and tools for identifying and characterizing these formats. This process enables the Develop Preservation Strategies and Standards function to issue recommendations and advice to deal with the incoming changes.

Carrying out these recommendations is the responsibility of the fourth function. The Develop Packaging Designs and Migration Plans function “develops new Information Package designs and detailed migration plans and prototypes, to implement Administration policies and directives.” This task will include development of new AIP designs, prototype software, test plans, community review plans and implementation plans for phasing in the new AIPs, and may call on expertise or resources from other functions. After proper testing and validation, the developed elements – plans, AIP designs and templates, software – will be sent as a package to be put into production.

2.2 Context of SPAR

SPAR (Scalable Preservation and Archiving Repository) is the BnF preservation system, compliant with the OAIS model. Its scope is to manage all entities that can be automated through modules corresponding to the OAIS entities.

2.2.1 Tracks and Channels

In SPAR, sets of documents to be ingested are processed by tracks and channels (sub-tracks), according to their nature (e.g., digitized books, audiovisual files, web archives, administrative records), their legal framework, and the way the BnF plans to manage their life cycle and apply preservation strategies. At the present time, SPAR ingests objects through six tracks: digitized documents and associated files, audiovisual objects, web legal deposit (ARC or WARC files), negotiated legal deposit (ebooks), administrative records, and third-party archiving (various kinds of files, from partners outside the institution); several others are in progress.

Every channel is managed by Service Level Agreements (SLAs), negotiated between the Producer and the Archive. They define the terms of ingest, preservation and dissemination (e.g., formats accepted, maximum size of packages, availability of service). Each SLA is transcribed in XML files that configure the system.

2.2.2 Reference Packages

SPAR is a self-documented system. It holds and preserves its own reference packages, ingested in a Reference track. These packages describe and identify every component of the preservation policy: formats, agents (software products, modules, processes, and humans), ontologies, classification systems, tracks and channels. SPAR uses them to document every process and, because most of them are machine-actionable, to perform automatic operations (e.g., checks, extractions, transformations). Thus, part of Representation Information and Preservation Description Information is preserved as well as the Data Objects, allowing reference to common information in every package through links and unique identifiers. This way, the verbosity of the manifests is reduced.

2.2.3 Actors

Many OAIS activities and functions cannot be fully automated and need human actors: administrators, preservation experts, developers, risk managers, collection managers and track managers. The Library has taken the measure of the challenge and is working on its organization (see [3], [4] and [5]).

Administrators are members of the IT department. They are responsible for deployment of new software versions, channels, requirements, etc. and, above all, for ensuring the system meets the SLAs in the daily production.

Preservation experts are members of many departments (e.g., from bibliographic information, IT, digitization departments, etc.); their expertise is functional or technical (on formats, storage, technical or bibliographic metadata, etc.). They are involved in standards elaboration. They are the core team that monitors the evolution of technology and the needs.

Track managers are members of departments who receive digital material to be preserved in SPAR: legal deposit, preservation, archiving mission, etc. They monitor a specific producer community and are responsible for ingest and preservation of documents belonging to their track.

2.3 The Preservation Planning Module: a Bottom-up Strategy

Up to now, development of SPAR was mainly concentrated on Ingest, Storage, Data management and Administration modules.

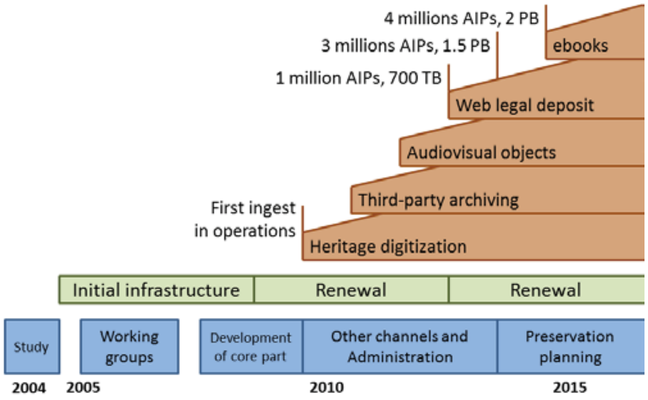


Figure 2. SPAR milestones

In 2014, the BnF decided to develop a module intended to fulfill the functions and activities of another OAIS entity: Preservation Planning.

2.3.1 Building Reference Packages

Formerly, the reference packages were discussed between project stakeholders then coded by developers in XML. Now they can be produced by the interface and modified at any time by a larger community of allowed users. Indeed, the new module is designed to foster collaborative work between administrators, track managers and preservation experts on the reference package elaboration.

In general, some of the benefits expected are:

- Greater speed and reactivity, involving common expertise throughout the library;
- Increased trustworthiness, thanks to a validation workflow involving more people; and
- Increased visibility of preservation activities.

2.3.2 *Documenting Decisions about Standards*

Whereas formats, channels and agents had always been preserved in SPAR, the Preservation planning module brings a whole new functionality. The entire decision process, from basic migration or characterization tests to preservation plans on a large scale, is now documented in reference packages ingested in SPAR.

Two major cases are foreseen:

1) Characterization. An upcoming file format has to be preserved in SPAR. What characterization tool will be used? What technical metadata is needed? See use case below, 4.1.
2) Migration. The new file has to be transformed when ingested into another format, preferably an open one. Which final format will be chosen? Which transformation tool will be used? See use case below, 4.2.

In the end, such tests will result in a new SLA with a new definition of ingest settings.

Four package types were defined to document this decision process. First, the Data Objects on which tests are carried out (initial "test data") and, in the case of a migration, the result of said tests (transformed "test data") are preserved. Secondly, characterization of initial and transformed data is preserved in "test metadata" packages referring to the Data Objects processed and to the used tool, described and ingested as an agent in the reference channel. One or more "test campaigns" are performed out of a significant number of tests leading to a decision that is implemented in a "preservation plan".

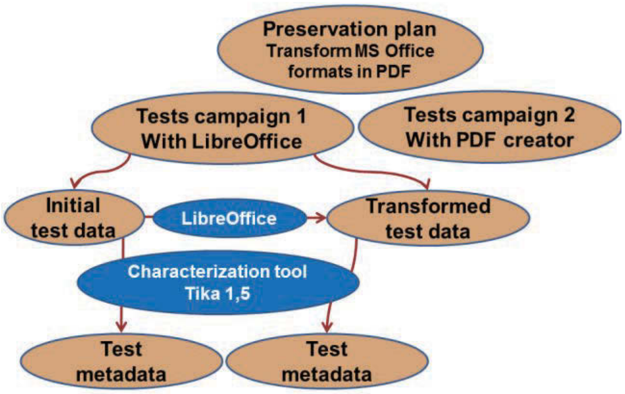

Preservation plan
Transform MS Office formats in PDF
Tests campaign 1 With LibreOffice
Tests campaign 2 With PDF creator
Initial test data
LibreOffice
Transformed test data
Characterization tool Tika 1,5
Test metadata
Test metadata

Figure 3. General organization of test packages

All this information is meant to be preserved in order to document the performed experiments, to give the material to allow the reproduction of these experiments, and to have a stable decision base in order to come back later and be able to reconsider such decisions.

3. IMPLEMENTATION CHOICES

3.1 A New Approval Workflow

In order to formalize the decision process, a validation circuit was organized, and several levels of authorization were defined according to it.

The following steps of the process were determined:

1) A specific need is submitted by track managers (e.g. a new file format to be preserved) or by preservation experts (e.g. a risk of obsolescence is identified for a specific format).
2) Tests are carried out locally by preservation experts to solve the issue (e.g. characterization and migration tools are run on sample documents).
3) If there has been a transformation, the transformed files are characterized by suitable software.
4) Sample documents and transformed documents are ingested in the Archive. Results of characterization are ingested as well.
5) A decision is taken about how to address the issue, given the tests results. Note that currently no method is defined to come up with the decision: anything such as mind maps, SWOT analysis or decision matrices can be used.
6) Preservation experts create new SLAs that take into account the new preservation strategy (the file format will be characterized or transformed into another format by a specific tool when ingested).
7) Programmers develop a technical solution (e.g. implementation of new characterization tools) and test it.
8) When ready, the technical solution is activated by the Administration.

3.2 General Architecture

The module architecture had to reflect the current organization of SPAR. At the same time the module was developed, a working group raised some important organizational issues about the role and attributions of every human agent related to SPAR. The Preservation planning module reflected these changes.

Managing several levels of authorization was a particularly important point in the module, as it gave direct capacity to SPAR's settings to agents out of the Administration module. Track managers have rights limited to definition of channels, whereas administrators can read and modify every type of package. Format experts have no rights on the channel packages but have a key role in the elaboration of formats and agent packages.

As the needs of the actors in SPAR are different and might be conflicting, multiple instances of the system have been installed.

- A validation platform is used by developers and the Quality Assurance (QA) team to build and test new versions of the software and to validate it.
- A sandbox platform is used by preservation experts and track managers to elaborate reference packages. This instance is also used for training and is regularly cleaned up.

- The production platform holds SPAR's current deployed version.

These instances are used in the approval workflow described above as follows:

1) The reference packages are elaborated collaboratively on the sandbox platform.
2) When ready, they are transferred to the QA team who tests them on the validation platform.
3) In the end, they are activated by the Administration and ingested in the production platform.

Going back to the Preservation Planning OAIS Entity, the SPAR implementation can be summarized as below:

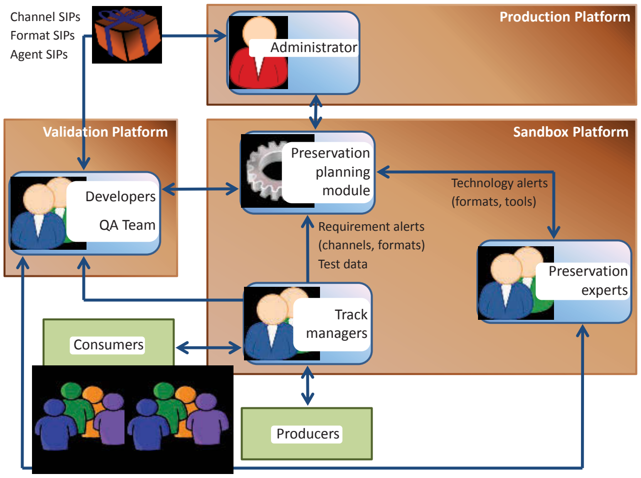


Figure 4. SPAR implementation of the Preservation Planning OAIS Entity

3.3 Ergonomics and Functionalities

Technically, the module aims to help building step by step a complete reference SIP and to transfer it into the repository. A previously ingested reference package can be updated with new requirements, thus creating a new version of this package.

Updating or creating a new package from an existing one is now easier: authorized users can ask for retrieval of an entire reference package, copy it, modify only the relevant information and ingest it again. The system delivers into a user-specific folder the manifest and the Data Objects contained in the requested package.

The interface usability was a challenge, as it was meant to bridge a gap between domain experts and IT staff. Vocabulary used in the interfaces had to be clear, consistent, precise and, preferably, agnostic about specific metadata terms.

In order to produce a complete machine-actionable reference package, different interfaces are displayed sequentially and cannot be accessed if the required information is not provided at each step. Compliance checks are carried out when moving from one interface to the next.

Common templates to several types of reference packages are defined to associate files, define events occurred before the package ingestion or enter its descriptive metadata.

The information provided by the user is recorded within the SIP manifest or within data files in which channel, software or format significant properties are stated. The preservation policy of the Reference track specifies that every version of the packages ingested by the track must be indefinitely preserved. In this way, one can always refer to requirements in effect at the time any event of package ingestion, preservation or dissemination occurred.

3.4 An Example: How to Create a Channel Reference Package?

Every channel in the SPAR repository has its own ingest requirements, preservation strategies and dissemination conditions.

The module allows track managers to create new channels and modify existing ones by updating the Service Level Agreements. This has immediate effects on packages ingest, preservation and dissemination.

The interfaces provide a set of information like patterns for descriptive metadata detection, possible transformations of input files, different files groups and formats allowed for each one, files minimum and maximum size, frequency of fixity checks, storage location, and documentation about the channel.

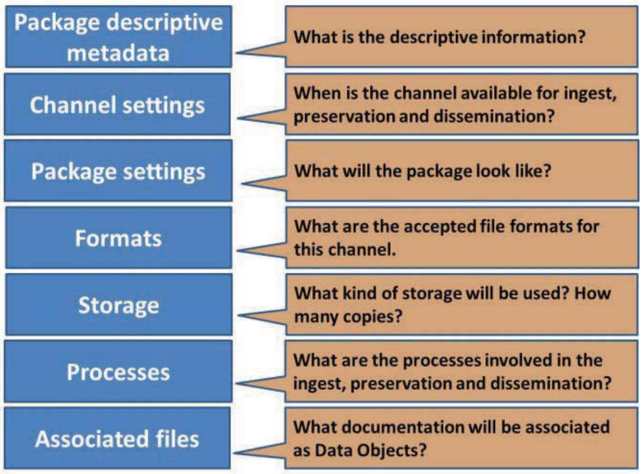

Package descriptive metadata
What is the descriptive information?
Channel settings
When is the channel available for ingest, preservation and dissemination?
Package settings
What will the package look like?
Formats
What are the accepted file formats for this channel.
Storage
What kind of storage will be used? How many copies?
Processes
What are the processes involved in the ingest, preservation and dissemination?
Associated files
What documentation will be associated as Data Objects?

Figure 5. Channel reference package elaboration sequence

At the end of the process, the reference package contains a METS manifest, a complete description of the channel, and the three SLAs (concerning packages ingest, preservation and dissemination) expressed in XML. Three associated Schematron files are used to check the manifests of every package submitted to the channel during its lifecycle.

4. SOME REAL USE CASES

4.1 A New Format for Heritage Digitization: JPEG 2000

The BnF's digitization program was primarily focused on producing images in uncompressed TIFF v6 format which is the preferential preservation format in this case. Due to the increasing volume of data (more than 1 PB), the switch to a compressed format was required by the track manager.

Thanks to the collaboration with other heritage institutions, the choice of the JPEG2000 format was appealing [6]. In order to determine which exact settings the Library should require for such a format, a set of sample TIFF images digitized from a vast diversity of material was assembled in a reference tests package. This package was then transformed using the kakadu tool with various settings and the results were compared in order to define the acceptable compression ratio in a similar fashion as described in [7]. In parallel, we use the jpylyzer tool [8] and an XSLT transformation to generate the corresponding test metadata package in the MIX format. We then had a way to ensure that the new images kept the significant properties of the images while taking less space.

Once the different settings were selected, the Digitized Program track manager was able to modify the reference package of its channel and the preservation system was able to ingest JPEG2000 files, characterized with the adequate tool.

4.2 Transforming Office Documents to PDF

In the course of elaborating the 'Administrative Records' track, a need for an additional preserved copy in PDF format of all the office documents became apparent.

Once again, a set of files were sampled from our production databases trying to target a large time period as well as to vary the versions of production software. As shown in [Figure 3], various tools to make the transformation were tried.

The question of finding a common format to represent the technical metadata led BnF to XMP [9], as the only one applicable for such a diversity of formats. The use of the Tika tool [10] to generate the test metadata packages provides a way to evaluate the well-formedness of the output as well as to compare the different outputs.

Currently, we have discovered that no tool is efficient enough to ensure a perfect transformation to PDF; such a conclusion reinforces the strategy of keeping both representations of the files (the original one and the transformed one) in the Archive.

5. CONCLUSION

As the module has been implemented recently, the BnF has little feedback from its potential users. Appropriation and community involvement will raise new issues and should be addressed in another paper in the years ahead.

However, using this module to elaborate in common SPAR standards has already shown good results, improving interaction between domain experts and IT members, and quality. But the module is one among other results of the BnF's will to involve more closely librarians in their digital collections preservation.

Finally, the module is likely to undergo evolution, as preservation planning encompasses many more aspects than only creating reference packages. Among them, being able to perform tests or migrations with tools stored in SPAR directly from the planning preservation module is foreseen.

6. ACKNOWLEDGMENTS

The authors would like to thank all the members of the SPAR team for their commitment and their invaluable contributions. The Open Preservation Foundation and all its members were paramount in sharing practices and maturation of ideas.

7. REFERENCES

[1] Becker, C., Kulovits, H., Rauber, A., and Hofman, H. 2008. Plato: a service-oriented decision support system for preservation planning. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries* (JCDL '08) (Pittsburgh, Pennsylvania, June 16-20, 2008). DOI= http://doi.acm.org/10.1145/1378889.1378954

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

[3] Bermès, E., and Fauduet, L. 2011. The Human Face of Digital Preservation: Organizational and Staff Challenges, and Initiatives at the Bibliothèque nationale de France. In *International Journal of Digital Curation*, vol. 6, no. 1, 226-237. http://www.ijdc.net/index.php/ijdc/article/view/175/244

[4] Derrot S., Fauduet L., Oury C., and Peyrard S. 2012. Preservation is Knowledge: A community-driven preservation approach. In *Proceedings of the 9th International Conference on Preservation of Digital Objects* (Toronto, Canada, October 2012). https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf

[5] Clatin, M., Fauduet, L., Oury, C., and Tramoni, J. P. 2014. Digital curators at work: analyzing emerging professional identities at the Bibliothèque nationale de France (BnF). In *IFLA World Library and Information Congress* (Lyon, France, August 2014). https://hal-bnf.archives-ouvertes.fr/hal-01098526

[6] Buckley, R. 2013. Using lossy jpeg 2000 compression for archival master files. http://www.digitizationguidelines.gov/still-image/documents/JP2LossyCompression.pdf

[7] Martin, S., and Macleod, M. 2013. Analysis of the variability in digitised images compared to the distortion introduced by compression. In *Proceedings of the 10th International Conference on Preservation of Digital Objects* (Lisbon, Portugal, 2013). http://purl.pt/24107/1/iPres2013_PDF/iPres2013-Proceedings.pdf

[8] Tarrant, D., and Van Der Knijff, J. 2012. Jpylyzer: Analysing jp2000 files with a community supported tool. In *Proceedings of the 9th International Conference on Preservation of Digital Objects* (Toronto, Canada, October 2012). http://eprints.soton.ac.uk/341992/1/iPres2012.pdf

[9] Extensible Metadata Platform (XMP), http://www.adobe.com/products/xmp.html, Last Access: 04/16/2015.

[10] Apache Tika. https://tika.apache.org/. Last Access: 04/16/2015.

Copyright and the Digitization of State Government Documents: A Preliminary Analysis

Brett Currier
UT Arlington Libraries
Central Library, Box 19497
Arlington, TX 76019-0497
+01 817-272-5127
brett.currier@uta.edu

Anne Gilliland
UNC Libraries
Davis Library, CB #3900
Chapel Hill, NC 27515-8890
+01 919 843 3256
agillila@unc.edu

David R. Hansen
UNC School of Law
160 Ridge Rd., CB #3850
Chapel Hill, NC 27599
+01 919 962 1605
drhansen@email.unc.edu

ABSTRACT

In this paper we explore the copyright status of state and local government documents and address some of the legal issues encountered when digitizing them.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows.

Keywords

Copyright; Government documents; Public records; Fair use; Digitization; State documents; North Carolina.

1. INTRODUCTION

State and local government documents make up an important and unique part of library and archives holdings. In addition to educational, scholarly, and research uses that motivate preservation and access to many other collections, libraries and archives have special care for state government records, which they maintain both to save the cultural history of their state and, especially for state institutions, to further promote open government. Copyright law poses a potentially significant obstacle to digitization and online access to state and local government documents, as it does for many other materials. Copyright Law grants owners six exclusive rights, covering reproduction, public distribution, public performance, public display, the right to control the preparation of derivative works, and a special right to control public performances of certain digital sound recordings.[24] Most important for digitization is the reproduction right, which has been interpreted broadly. Unless an exception such as fair use applies, mere reproduction into a dark archive--even with no associated distribution or display to the public-- would implicate the copyright owner's rights.

This is important because digital duplication and reproduction is becoming an archival standard as analog items are shifted to and preserved in digital formats and digital items are preserved in their original formats. One strategy for preservation of analog materials includes shifting the material into a digital format.[17] But again, mere reproduction of an analog item in a digital format implicates the copyright owner's rights. Archival standards of preservation of digital items may include web hosting by a memory institution instead of the original content creator,[10] creating a copy through routine backups,[11] or creating duplicate copies of the items on backup servers.[1] All three of those implicate the copyright owners' reproduction right. Thus, as decisions are made regarding work flow and preservation strategies of state and local government documents, an analysis of relevant copyright law should be included as those workflows and preservation strategies are created.

Copyright law affects preservation of state government records as it does preservation of many other types of works. But in terms of laws governing their reuse, government records are unique because their creation and use is governed not just by federal copyright law, but by state and local laws, such as public records acts, that provide additional opportunities for use. And even in cases where state public records laws are unclear, several common characteristics of state and local government records collections make them more amenable to use because they are either non-copyrightable subject matter (e.g., factual data sets), in the public domain, or usable under fair use or other copyright limitations. This paper provides a preliminary analysis of relevant copyright law for documents produced by the state and local government and collected by libraries and archives, emphasizing uses in one state—North Carolina—as an example. It includes ideas about how librarians and archivists can use that law to further digitization efforts and to provide access to these materials.

2. COPYRIGHTABILITY OF STATE AND LOCAL GOVERNMENT DOCUMENTS

Copyright applies, in general, to any "original work of authorship fixed in a tangible medium of expression."[21] Within that definition of protectable works are several significant limitations. First, the work must be original, meaning that an exact reproduction (for example, a digital surrogate of print work) does not itself receive protection.[2] Second is that protection extends only to a work of authorship. As a matter of statute and the U.S. Constitution, courts have held that copyright can only extend to creative works; publications that merely report unadorned facts are not protectable.[6] And third, the work must be fixed—typically not something in question with government publications.

In addition to those general limitations, Congress has created a categorical exclusion for some government works. Under Section 105 of the U.S. Copyright Act, federal government works are not protectable in the U.S,[23] nor are works ". . . prepared by a [federal] government employee as part of the employee's official duties."[28] While this rule is limited (Congress excepted from this rule, for example, many works created by government contractors),[28] most federal government documents are not protected by federal copyright law.

When it enacted Section 105 of the Copyright Act, Congress considered and rejected applying the same exclusion from protection to state governments and to works of foreign nations.[7] Thus, works of state and local governments, as a category, are not excluded from federal copyright law protection.

However, large numbers of state and local government works do not receive copyright protection because they are not copyrightable subject matter under the more general exclusions—that is, because they are not sufficiently original, or because they contain only facts and no creative expression. Courts have most notably dismissed claims of copyright material if an individual had attempted to copyright primary law or edicts of government, such as case law, statutes, state regulations, or municipal codes.[29] Along with Constitutional concerns with applying copyright to those materials, courts have concluded that those edicts of government are facts, and thus cannot be protected.[27] While some of these materials have been registered with the U.S. Copyright Office, that alone does not establish copyrightability [5]

State and local government also produce an array of data sets and factual publications, covering everything from vital statistics to agriculture. Likewise, states and local governments are now large producers of geographic data, created for a wide variety of GIS mapping applications. The selection and arrangement of those facts, if creative and not merely dictated by convention, may, as a whole, be protectable as a copyrighted compilation.[22] But the underlying data is free to be reused because it is not a work of authorship, but fact. What protection there is in the compilation is limited to the creative elements of the compilation.

However, not all or even most state and local government publications would be considered non-copyrightable subject matter. Many state and local publications are highly creative. One good example of highly creative work by a state agency is the North Carolina Film Board (NCFB), the first state sponsored documentary film division of any state government.[12] The NCBF created at least 30 documentary films during its existence from 1962-1965.[15] Works need not be nearly so creative, however, to qualify for copyright protection. Even relatively straightforward reporting and summation—for example, in an environment impact report—could qualify for protection. Thus libraries and archives must look to other legal provisions to make uses of those works.

3. STATE LAW EXCLUSIONS FROM FEDERAL COPYRIGHT PROTECTION

In addition to federal limitations on copyright's application, states have self-imposed rules that limit the application of copyright to government works. Typically, these declarations are contained in states' public records laws that seek to provide transparency and access to state and local government activities. There is no uniform state public records law, however, and in the rare instances in which state governments and courts have weighed in on the interaction between public records laws and copyright, results have varied. In New York, for example, the U.S. Court of Appeals for the Second Circuit held that New York's public records law "does not prohibit a state agency from placing restrictions on how a record, if it were copyrighted, could be subsequently distributed."[5] Similarly, South Carolina has stated that so long as the public records are sufficiently creative and original, nothing in the state public records law would prohibit the state or local governments from exerting copyright protection.[16] Florida and California courts, however, have concluded that end-user restrictions imposed by copyright would be incompatible with the purpose of their public records acts, to provide public transparency.[13]

Many states have no clear statement about whether their state's records are subject to copyright protection. North Carolina's Public Records Act, for example, declares that public records—defined broadly to encompass all documents produced in connection with public business regardless of format—are "the property of the people."[14] Federal law sometimes refers to the "public domain" (though it is not statutorily defined), but does not anywhere use the phrase "property of the people." Because of the lack of a definitive statement about the copyright status of North Carolina state documents, librarians and archivists are left to make their own conclusions. The North Carolina State Library has done just that, putting users on notice that, for public records, it asserts that those works are in the public domain and eligible for reuse.[19]

4. THE PUBLIC DOMAIN

Modern copyright protection lasts longer now than at any time in the past. In the United States, the standard term of protection extends for the life of the author plus an additional seventy years.[8] Life plus seventy years is also a common international norm. For U.S. institutions, however, requirements under prior law that copyright owners renew their copyright term via registration and provide a copyright notice means that many older works, both government publications and private works alike, have entered the public domain.

Public domain analysis can be complex, though some clear rules apply. For example, state government works published in the United States before 1923 are clearly in the public domain. Likewise, state government works published in 1989 or later (and not excluded from protection because of one of the earlier-discussed exceptions) are protected by copyright law. Distinguishing between "published" and "unpublished" state government documents—a necessary inquiry to determine public domain status, as different rules apply to those two categories—can requires significant investigation into how the document was acquired and how it was originally released. There are a significant number of state government works published between 1923 and 1989 that, in order to receive copyright protection, must have complied with federal formalities. To determine the copyright status of those works requires significant research into copyright office records and the work itself. Several efforts to develop a methodology and workflow for this analysis are in development, notably through the IMLS-funded Copyright Review Management System (CRMS) at the University of Michigan Libraries.[26]

Libraries and archives that do undertake this analysis using processes like those developed by CRMS are likely to find that many state publications, especially those never offered for sale and distributed freely, were published without required notices or were never renewed, causing the work to enter the public domain and thus available for reproduction and other reuses.

5. FAIR USE AND OTHER EXCEPTIONS

Finally, U.S. libraries and archives have access to fair use and other copyright exceptions that allow for certain types of uses of state government works even when the work is protected by copyright law.

The fair use doctrine, created by courts and now codified in the U.S. Copyright Act, asks users and the courts to consider several factors, four of which are explicitly identified in the statute: (1) the purpose and character of the use, (2) the nature of the work, (3) the amount and substantiality of the portion used, and finally, (4) the effect of the use upon the marketplace.[25] Courts must weigh those factors together, in light of the purposes of copyright law.[3] Although there are few reported cases challenging library and archive uses, at least one court has now weighed in on fair use as applied to library and archive uses. In that case, *Authors Guild v. HathiTrust,* the Second Circuit Court of Appeals found that the HathiTrust Digital Library's digitization for purposes of preservation and search were fair use.[1]

Fair use is highly fact dependent, and so it is helpful to analyze its applications to common scenario: a library's digital preservation and full-text online access to a technical report that was published by a state agency and distributed in print free of charge. Applying the fair use factors, factor one (the purpose and character of the use), would likely weigh in favor of a finding of fair use because the purpose of the reproduction and distribution is to promote access and openness to government and to preserve them for the future. Factor two, the nature of the work, would likely also weigh in favor of a finding of fair use; the work was distributed to the public for free, and copyright law was unlikely to have motivate this work's creation.[4] In addition, as a technical report its contents are likely factual. While still sufficiently creative to trigger copyright protection, factual reporting of this nature is likely to be favorably viewed for use under the second fair use factor. Factor three, the amount and substantiality of the portion used, would if anything weigh against a finding of fair use because the entire work is digitized, but courts have often found even this to be unavailing if the amount taken is appropriate within the context of the other factors. Finally, factor four, the effect of the use upon the marketplace, would likely weigh in favor of a finding of fair use. Most copyrighted materials produced by state and local governments are information providing and not revenue producing materials. For many materials produced by state and local governments, the government is only able to recoup the actual cost of reproduction.[14] Some materials are more similar to traditional revenue producing models like The North Carolina Museum of Art, a division of the North Carolina Department of Cultural Resources, which produces exhibition catalogues. A finding that the fourth factor favors fair use would be more likely once the work is no longer published, remains out of print, and is no longer for sale from the copyright owner.

All in all, it is likely that fair use supplies a good rationale for the digitization of many state documents, particularly for those that are highly factual and not commercialized. In addition, other copyright exceptions, such as the exceptions that allow libraries to make reproductions for preservation, may be relevant.

6. RISK MANAGEMENT

The actual risk of litigation and the risk of losing any lawsuit brought against a digitizer of state documents is unknown. Certainly most cultural institutions that might start such projects are ill equipped, both financially and temperamentally, to engage in an extended defense of their practices, even if they are likely to win in the end. There is ample precedent to argue the public domain status of state legal codes of all kinds, but there are little or no cases on point with regard to the copyright status of other state documents.[9] At the same time, there are costs to be paid for inaction, both by losing information if the historical record is not preserved and also in depriving the public of easy access to its government's publication. In some cases, there is a concern that keeping governmental publication and records can be a form of censorship.

Generally, the strategies of risk management used in many libraries' large-scale digitization projects will also apply for digitization of state government information. These strategies include identifying work that is likely not to have passed into the public domain and then further making attempts to identify and contact rights holder. Essentially, the idea is to focus on clearing rights in instances where identifying a rights holder is likely to be both possible and prudent. Other material is digitized and posted online with an invitation for rights holders and others to get in touch, either in order to provide information on copyright status or contribute more background and identification for the material.[20]

CONCLUSION

The copyright status of state and local government documents is not a settled issue, and thoughtful consideration of a variety of factors must precede plans to digitize this material. The first is whether the material is copyrightable under federal law. Statistical information, for example, may have no copyright as factual material. In addition, state law may yield clues to the copyright status of state documents, either directly or indirectly. In North Carolina, public records laws, federal copyright law and public policy considerations lend some credence to the idea that many state documents are in the public domain and should be freely reproduced and distributed for preservation and access by the public. Finally, fair use and copyright exceptions for libraries may provide a rationale for digitization.

REFERENCES

[1] Authors Guild, Inc. v. HathiTrust, 755 F. 3d 87 (2nd Circ., 2014)

[2] *See* Bridgeman Art Library, Ltd. v. Corel Corp., 25 F. Supp. 2d 421, 426-27 (S.D.N.Y. 1998).

[3] Campbell v. Acuff-Rose Music, 510 U.S. 569 (1994).

[4] See, e.g., Connecticut State Library. *Federal Depository Library Program*, 2015 http://ctstatelibrary.org/access-services/fdlp/.

[5] County of Suffolk v. First American Real Estate Solutions, 261 F.3d 179 (2d Cir. 2001).

[6] *See* Feist Publications, Inc. v. Rural Telephone Service Co., 499 US 340 (1991).

[7] *See* H.R. Rep. 94-1476.

[8] Hirtle, P. Jan. 1, 2015. Copyright Term and the Public Domain in the United States. https://copyright.cornell.edu/resources/publicdomain.cfm

[9] *See* Li, V. 2014. Who owns the law: technology reignites the war over just how public documents should be. *ABA J* (June 1, 2014).

[10] Library of Congress. (n.d.) Digital Preservation. Retrieved July 8, 215, from http://www.digitalpreservation.gov/

[11] Lynch, C. (2003). Institutional repositories: essential Infrastructure for scholarship in the digital age. *portal: Libraries and the Academy, 3*, 2 (April 2003), 327-336.

[12] Mazzochi, J. 2006. North Carolina Film Board. In *NCPedia*. State Library of North Carolina. http://ncpedia.org/north-carolina-film-board

[13] Microdecisions, Inc. v. Skinner, 889 So. 2d 871 (Fla. Dist. Ct. App. 2004); County of Santa Clara v. Cal. First Amendment Coal, 89 Cal. Rptr. 3d 374 (Cal. Ct. App. 2009).

[14] N.C. Gen. Stat. 132-1(b) (2014)

[15] Oettinger, E. The North Carolina Film Board: A Unique Program in Documentary and Educational Film Making. *The Journal of the Society of Cinematologists*, 4, (1964/65), 55-65.

[16] Seago v. Horry County, 663 S.E.2d 38 (S.C. 2008)

[17] Smithsonian Institution Archives. (n.d.). *Collections Care*. http://siarchives.si.edu/services/collections-care.

[18] *See, e.g*., State of Minnesota: Supreme Court. *Minnesota Reports: Vol. 151: Cases Argued and Determined in the Supreme Court of the State of Minnesota between December 23, 1921 and April 7, 1922*. Copyright Registration 698314, Issued February 13, 1923; State of Missouri: Supreme Court. *Missouri Reports: Reports of Cases Argued and Determined in the Supreme Court of the State of Missouri between July 11, 1921 and December 19, 1921: v. 289 and 290.* Copyright Registrations 696577 and 696578, Issued January 25, 1923 and February 8, 1923. Some later registrations acknowledge that there is no copyright in the underlying primary state law, but not all. *Compare* Sandra Parker Hoffpauir. *Health records and Alabama law*. Copyright Registration TX0005783662, Issued June 14, 2001 *with* Michie, a division of Reed Elsevier, Inc. *Kentucky revised statutes annotated : vol. 8A, ch. 186-190 : certified version : containing statute laws of the commonwealth of a general and permanent nature in effect as of October 1, 1997 / compiled under the authority and supervision of the Legislative Research Commission.* Copyright Registration TX0004635445, Issued December 22, 1997.

[19] State Library of North Carolina. North Carolina digital collections [rights statement] http://digital.ncdcr.gov/cdm/ref/collection/p249901coll22/id/63754); see also *State Copyright*, Copyright at Harvard, http://copyright.lib.harvard.edu/states/ (last visited Nov. 20, 2015) (compiling statements on state copyright assertion from all fifty states).

[20] See Stobo, V, Deazley, R. and Anderson, I.G. 2013. Copyright & risk: scoping the Wellcome Digital Library Project. *CREATe Working Paper* 10 (2013); Smith, K.L. 2012. Copyright risk management: principles and strategies for large-scale digitization projects in special collections. *Research Library Issues. 279, (*June, 2012) 17-23. http://publications.arl.org/rli279/17. For good examples of this kind of risk management in practice, see New York Public Library, "Reply Comments to Orphan Works and Mass Digitization: Notice of Inquiry" (77 F.R. 64,555, Docket No. 2012-12), March 16, 2013, http://copyright.gov/orphan/comments/noi_11302012/New-York-Public-Library.pdf (describing challenges in finding owners and clearing rights for works relating to the 1939-1940 World's Fair) and UCLA Libraries. Library Special Collections Risk Assessment Guidelines, https://www.library.ucla.edu/sites/default/files/Guidelines_RiskAssessment.pdf (defining procedures for developing a workflow for establishing a matrix of risk)

[21] 17 U.S.C. § 102 (2012).

[22] 17 U.S.C. § 103 (2012).

[23] 17 USC § 105 (2012).

[24] 17 U.S.C. § 106 (2012).

[25] 17 U.S.C. § 107 (2012).

[26] University of Michigan. 2015. *Copyright Review Management System*. http://www.lib.umich.edu/copyright-review-management-system-imls-national-leadership-grant

[27] Veeck v. Southern Bldg. Code Congress Intern., Inc., 293 F.3d 791 (5th Cir., 2002).

[28] Walton v. U.S., 80 Fed.Cl. 251, 271 n. 30 (United States Court of Federal Claims, 2008)

[29] Wheaton v. Peters, 33 U.S. 591 (1834)

Project Chrysalis –Transforming the Digital Business of the National Archives of Australia

Zoe D'Arcy
National Archives of Australia
Queen Victoria Terrace
Parkes, ACT 2600
612 6212 3606
zoe.darcy@naa.gov.au

ABSTRACT

The role of the National Archives of Australia is to promote the creation, management and preservation of authentic, reliable and usable Commonwealth government records and enable ongoing public access to the archival resources of the Commonwealth.

Records that are created by Commonwealth government agencies and transferred to the National Archives are, of course, predominately digital. Digital records bring a range of challenges, but they also potentially present new opportunities in the way archives can conduct their business. This paper outlines a project currently underway at the National Archives, named Project Chrysalis, which is an end-to-end business system that aims to transform the way in which the Archives does its digital business.

Project Chrysalis represents not just a technical solution, but also significant business change for the National Archives. However, if implemented successfully, the project should enable the Archives to sustainably harvest, preserve and provide access to digital records in the information age.

General Terms

Institutional opportunities and challenges; Technical opportunities and challenges; Innovative practice; Metadata; Automation of Processes; Machine Learning

Keywords

Government; digital records; business system; metadata; automation; machine learning; change.

1. CURRENT ARCHIVES AND COMMONWEALTH GOVERNMENT ENVIRONMENT

The National Archives has been actively in the digital space from the late 1990s. The Archives provides information policy, advice and training to Commonwealth government agencies so that digital records are created and managed appropriately. The Archives transfers, preserves and manages both digital and analogue records of permanent value (RNA).

The Archives' services to the public are also predominantly digital. It digitises analogue records already held in its collections, and in the last financial year, 99% of collection access to paper records took place online rather than in an Archives' Reading Room.

Commonwealth government agencies are creating digital records. In 2011, the Australian Government Digital Transition Policy was approved by Cabinet. Under this policy, Commonwealth government agency records that are created digitally after 2015 must be kept in a digital format and those identified as RNA must be transferred to the Archives in digital format.

As a consequence of this policy:

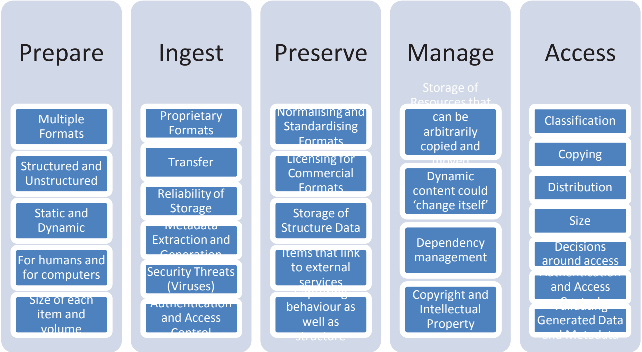
- Many Commonwealth agencies are managing their digital records digitally e.g. with an Electronic Document Record Management System (EDRMS)
- Many Commonwealth agencies are digitising their physical records.
- It is estimated that digital transfers to the Archives will grow to 32 TB/annum by 2020.
- To meet this expected increase a review of digital processes and systems was conducted in 2012. The review concluded that the Archives needed to:
- Develop our business capabilities in order to sustainably harvest, preserve and provide access to digital records
- Increase our capacity to provide online access to both digital and analogue collections
- Create a rich metadata structure that allows for enhanced search and discovery for both agency and public clients.

2. CHALLENGES AROUND DIGITAL

Managing digital records brings a range of challenges, and like many archives round the world, the National Archives must develop its business capabilities in order to sustainably harvest, preserve and provide access to its born-digital collection, as well as increase its capacity to provide online access to its analogue collection.

The primary challenges of digital records for Government Agencies and the Archives through their life-cycle are outlined in this table.

Diagram 1. Primary challenges of digital records throughout their life-cycle


Prepare
Multiple Formats
Structured and Unstructured
Static and Dynamic
For humans and for computers
Size of each item and volume
Ingest
Proprietary Formats
Transfer
Reliability of Storage
Metadata Extraction and Generation
Security Threats (Viruses)
Authentication and Access Control
Preserve
Normalising and Standardising Formats
Licensing for Commercial Formats
Storage of Structure Data
Items that link to external services
behaviour as well as structure
Manage
Storage of Resources that can be arbitrarily copied and
Dynamic content could 'change itself'
Dependency management
Copyright and Intellectual Property
Access
Classification
Copying
Distribution
Size
Decisions around access
Authentication and Access Control
Generated Data and Metadata

However, the richness of data in digital records also offer technical opportunities, and it is those that Project Chrysalis seeks to exploit.

3. MEETING THE CHALLENGES – PROJECT CHRYSALIS VISION

The National Archives' digital business solution Project Chrysalis has been created to meet these challenges. Based on the Open Archival Information System (OAIS), it aims to:

- Provide online access to the records as soon as possible, to clients who are anywhere, at the time the records are required, and in a format and on platforms that meets their requirements
- Preserve and manage the Commonwealth's digital records ensuring their long term integrity and authenticity
- Enable cost effective and efficient sentencing of Agency digital records and their transfer to the Archives

This envisioned technical solution is not a single system but consists of modular, integrated components that provide a scalable, extensible platform for digital business.

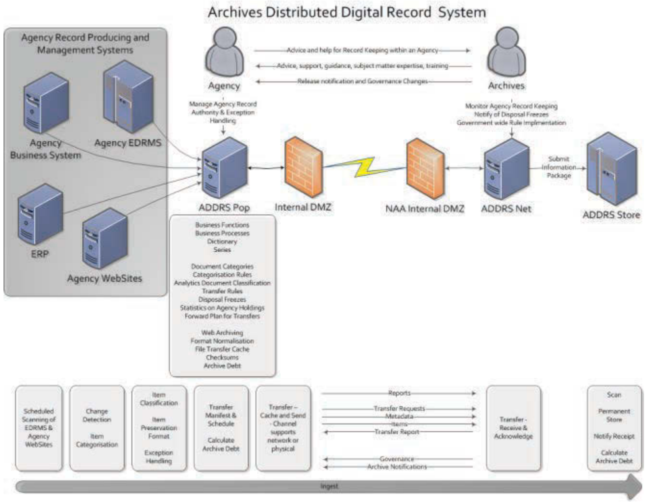
4. DESIRED OUTCOMES

4.1 From Government Agencies to Archives

In 2014 the National Archives engaged three Solution Architects were engaged to design a Digital Business Architecture Blueprint. The Blueprint details the end state for the Business, Technical and Information architectures for the solution. The implementation of the blueprint is called Project Chrysalis.

This diagram shows the proposed end-state for the transfer of digital records the Archives.

Diagram 1: Proposed end-state for the transfer of digital records to the National Archives


Archives Distributed Digital Record System
Agency Record Producing and Management Systems
Agency
Archives
Advice and help for Record Keeping within an Agency
Advice, support, guidance, subject matter expertise, training
Release notification and Governance Changes
Agency Business System
Agency EDRMS
ERP
Agency WebSites
Manage Agency Record Authority & Exception Handling
Monitor Agency Record Keeping Notify of Disposal Freezes Government wide Rule Implmentation
ADDRS Pop
Internal DMZ
NAA Internal DMZ
ADDRS Net
Submit Information Package
ADDRS Store
Business Functions
Business Processes
Dictionary
Series
Document Categories
Categorisation Rules
Analytics Document Classification
Transfer Rules
Disposal Freezes
Statistics on Agency Holdings
Forward Plan for Transfers
Web Archiving
Format Normalisation
File Transfer Cache
Checksums
Archive Debt
Scheduled Scanning of EDRMS & Agency WebSites
Change Detection
Item Categorisation
Item Classification
Item Preservation Format
Exception Handling
Transfer Manifest & Schedule
Calculate Archive Debt
Transfer – Cache and Send - Channel supports network or physical
Reports
Transfer Requests
Metadata
Items
Transfer Report
Governance
Archive Notifications
Transfer - Receive & Acknowledge
Scan
Permanent Store
Notify Receipt
Calculate Archive Debt
Ingest

Agencies will have their record producing and management systems connected to or integrated with the Archives Distributed Digital Record System (ADDRS).

The ADDRS point of presence (POP) tool will be a digital 'records authority' where it will hold information about an agency's functions, systems and classification/categorisation rules, etc. This tool will enable the records that are RNA to be identified, exported and/or harvested, batched and transferred to the Archives via an automated process over the most appropriate channel.

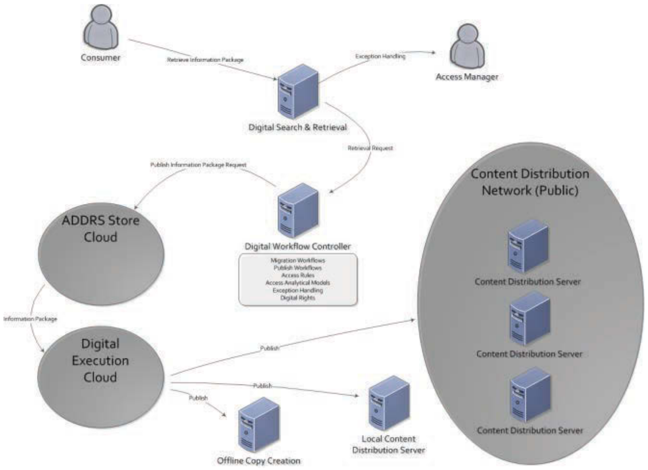
The workflow show the processes that will occur. Starting on the left, systems will be scanned on a scheduled basis, records and metadata will be exported and/or harvested from agency systems, checked, classified, checksum applied, converted to a preservation format and transferred to the Archives.

The records will then be ingested into the Archives, quality assured and stored for further processing within Archives.

4.2 From Archives to the Consumer

This diagram shows the desired outcomes for how consumers will access records from the Archives digital and physical collections. Consumers can be Archives staff, Agency and Public clients.

Diagram 2. Proposed end state for the delivery of digital records to consumers


Consumer
Retrieve Information Package
Exception Handling
Access Manager
Digital Search & Retrieval
Retrieval Request
Publish Information Package Request
Digital Workflow Controller
Migration Workflows
Publish Workflows
Access Rules
Access Analytical Models
Exception Handling
Digital Rights
ADDRS Store Cloud
Information Package
Digital Execution Cloud
Publish
Publish
Publish
Offline Copy Creation
Local Content Distribution Server
Content Distribution Network (Public)
Content Distribution Server
Content Distribution Server
Content Distribution Server

All records ie paper-based (not digitised) records, paper-based digitised records, born digital records, AV records etc. through a single, federated search and discovery function:

- Clients will be able to search across multiple Archive repositories, potentially including external agency data sources and secure Cloud environments.
- The retrieved records will be published to a delivery platform most appropriate for the client, regardless of the software on their device. For example, a digital file could be published and made accessible via a web download – potentially using a third party service such as Google Drive.

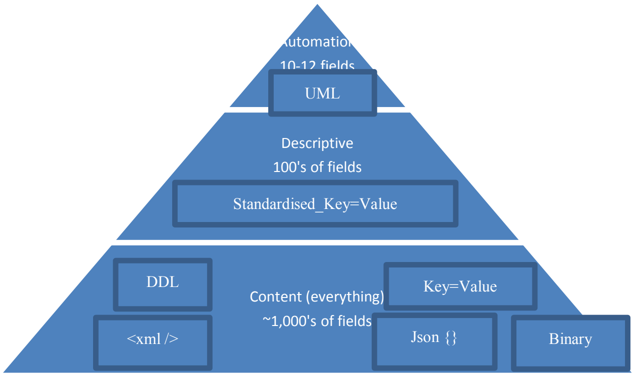
4.3 Designing for Complexity and Scale

Project Chrysalis is being prototyped using records from EDRMS systems; however, it is being designed from the outset to be able to deal with complexity and scale that the National Archives is expecting of digital records. There are three key areas that we're hoping will enable us to do so sustainably: metadata, automation of processes and use of machine learning.

4.3.1 Metadata

The National Archives actively promotes the use of two metadata schemas for use of Australian government agencies – AGRKMS for government digital records, and AGLS for government websites – both are based on Dublin Core. Technically, however, the Archives has to support a logical metadata model that allows the ingest records that conform to the much wider range of metadata schemas that are in common use amongst government agencies. We have to store those records; manage and automate business processes that enhance a record or move it from one state to another; and also have a searchable index of the records.

Diagram 3: Metadata Pyramid



Automation Metadata – this data is required to support the management component, and will be managed in a Relation Database Management System. This answers questions like: What state is the information package currently in? Why did the information package change state? Who currently owns it? What format is it in? What is the security level? This information must be accurate and unambiguous as it will be used by the computer system to orchestrate and perform transactions on the information package itself.

Description Metadata – this is the data that is required by the index in a full-text search engine to allow the information package to be found and retrieved from the storage. This may include discovered/derived information, descriptions, annotations, transcriptions, summaries, extra context and textual content.

Content – this is the actual information package. It is the information package that is stored within the object store. It must be able to be retrieved, based on a unique identifier, and be in a format that can be read by the end user. It may also include additional information that has been added before, during and after transfer of the Information Package to the Archives.

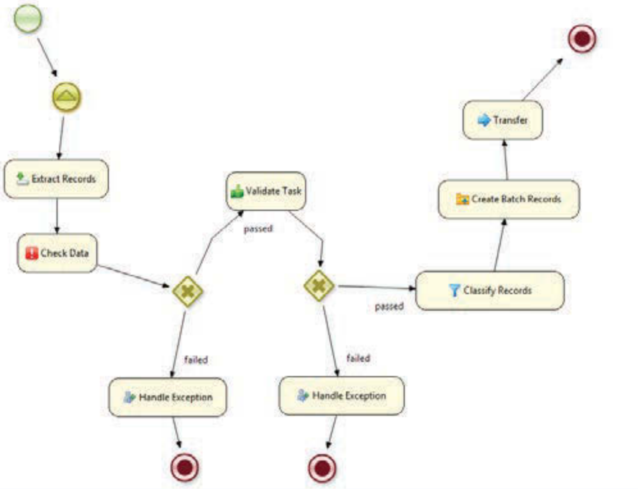
These layers are not expected to be distinct or static. As business processes change, it is possible that new or different automation metadata will be required. As information packages are described or new types ingested then new descriptive metadata will be required. And, of course, as new applications and technologies are used by our client then new content will be coming in to the Archives.

4.3.2 Automation of Processes

One of the features of Project Chrysalis architecture is the use of business rules to automate as many of the National Archives workflow processes as possible. While human decision-making will always be completely necessary for the Archives' technical solution to work, automation of certain processes will allow scalability and sustainability.

The diagram below shows an automated Records Extraction process - the extraction of the records and metadata from a Commonwealth government agency system to be loaded to a Submission Information Package (SIP) and transferred from the agency to the Archives.

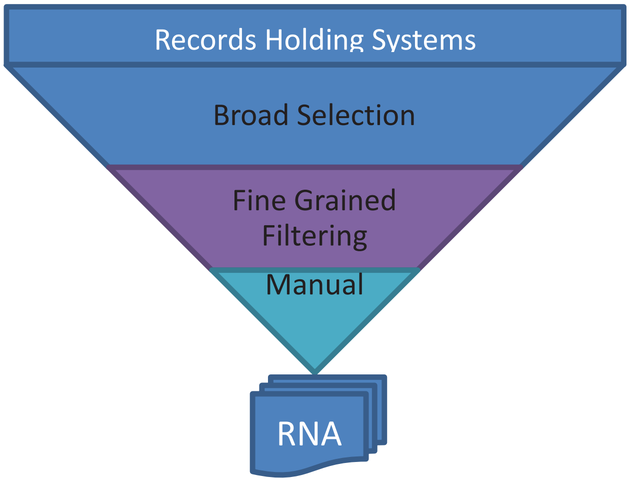
Diagram 4. Example of business rules that will enable automation – the records extraction process



4.3.3 Use of Machine Learning

At several points in the workflow processes, Project Chrysalis looks to using machine learning tools for assistance with the scale of digital records. For instance, we know that one of the key challenges for our government agency clients is that the National Archives does not want to take all of their records – only those that are classed as 'Retain as National Archives' (RNA). The current process of records selection is very manual. We have prototyped a tool for use by staff within those agencies to search across records holding systems for RNA records, and begin training the tool which records do and do not fall into that category, so that they can quickly be assisted in this classification process.

Diagram 5: Selection of Records for Transfer to the National Archives of Australia



5. GETTING THERE

5.1 Technical Change

In these diagrams everything works beautifully. However, they also cover some very ambitious ideas which may take some time to achieve. So how is the National Archives planning to get to the proposed end state?

The answer is a staged, iterative approach. Each iteration builds on the previous one, and each is expected to take two years. Within each iteration there are evaluations points to assess our approach and determine what is working and what is not. In brief, they are as follows:

- **Iteration 1** is a Proof of Concept with the aims to test and validate the architectural concepts and technologies identified for the Archives digital business solution and build a prototype of the Archives digital business system.
- **Iteration 2** extends the trial to transferring data from selected partner agencies to test scalability & automation of transfer & storage. It aims to do this without increasing costs for either the Agencies or the Archives through automation.
- **Iteration 3** builds on Iterations #1 and #2 and extends the trial to all Archives' consumers, including agency clients. Iteration 3, tests functionality that allows for the efficient finding, viewing and retrieval of records. These functions will need to scale to support both ad-hoc consumers, all the way up to 'big data' consumers who require terabytes of data.
- **Iteration 4** trials the extraction of agency data directly from the agency and/or gateways and plans for full production. It will achieve the vision of Project Chrysalis.
- **Iteration 5** is business as usual. The solution is in place to transfer, preserve, manage and provide access to digital records on an ongoing basis.

The National Archives is currently in Iteration 1. In July 2015 we successfully finished the proof of concept. Using the Archives' own internal records from its EDRMS, we proved that it is technically possible to develop a suite of software to assist record keepers in the digital age by:

- At the Client Agency: digitally selecting and "packing" records stored in a client agency's EDRMS and transmitting that package to the National Archives of Australia for management.
- At the Archives: receiving, "unpacking", storing, preserving and digital records received from client agencies in a secure environment.
- In the world, online: providing Agencies with secure private access to their transferred records, and providing the Australian public with greatly enhanced discovery and interaction opportunities through a federated and faceted discovery experience.

Lessons learnt from the build of the proof of concept will inform decisions, processes and planning required for the development of the prototype. The end-to-end prototype is scheduled to be completed by June 2016.

5.2 Organisational Change

In order for the technical solution to be successful, the Archives will also have to change how it performs its business and the services it offers. The transformation required within the organisation will include re-imagining process that work for analogue records - re-engineering digital business processes so they are:

- reliable, robust and sustainable
- clear definitions how the business operates, who owns the processes and how this links into the overall operation of the Archives' business
- flexible enough to implement rapid change and meet client demands whilst maintaining data integrity
- able to deliver productivity improvements
- Information management policy, advice and standards will also need to change, so that they can:
- increase the Archives' ability to manage its digital business end-to-end
- facilitates government agencies in managing their digital records without increasing costs
- enables Archives and clients to utilise data from multiple sources to provide meaningful content and related information
- make explicit the cost of preserving, storing and a providing ongoing access to digital records

6. BENEFITS OF DIGITAL TRANSFORMATION

Transformation of business is never easy, but we believe that the benefits of successful implementation of Project Chrysalis will see some real benefits for the National Archives and its clients – both government agencies and public researchers.

- It will be easier for agencies to transfer digital records to the Archives, as there will be standardised and automated transfer (export/harvesting) for most agencies
- Support for distributed custody and access to digital records that cannot be easily transferred
- Metadata standards that can be built upon and utilised to drive and enhance workflow automation and facilitate finding, viewing and retrieval of records
- Ability to retrieve digitised paper records due to content indexing and search functions
- Support for consistent digital access to the Archives' collection via multiple channels

We also think that the technical solutions should provide some cost savings via:

- Increased efficiencies and higher productivity (e.g. more automated processes);
- Better/more reliable reporting for Archives and Agencies, as it will be more in real time
- Lower cost of system ownership (e.g. reduced maintenance/support effort, lower support costs)
- Flexibility & adaptability to handle changing business requirements without necessitating development and support of new systems

Project Chrysalis is in its early days. It has been designed to be an end-to-end business system to enable the National Archives to manage digital records in a way that takes full advantage of the benefits of digital information. If implemented successfully, the project should enable the Archives to sustainably harvest, preserve and provide access to digital records in the information age.

Benchmarks for Digital Preservation tools

Kresimir Duretec [1], Artur Kulmukhametov[1], Andreas Rauber[1] and Christoph Becker[1,2]

[1]Vienna University of Technology, Austria

[2]University of Toronto, Canada

ABSTRACT

Creation and improvement of tools for digital preservation is a difficult task without an established way to assess any progress in their quality. This happens due to low presence of solid evidence and a lack of accessible approaches to create such evidence. Software benchmarking, as an empirical method, is used in various fields to provide objective evidence about the quality of software tools. However, the digital preservation field is still missing a proper adoption of that method. This paper establishes a theory of benchmarking of tools in digital preservation as a solid method for gathering and sharing the evidence needed to achieve widespread improvements in tool quality. To this end, we discuss and synthesize literature and experience on the theory and practice of benchmarking as a method and define a conceptual framework for benchmarks in digital preservation. Four benchmarks that address different digital preservation scenarios are presented. We compare existing reports on tool evaluation and how they address the main components of benchmarking, and we discuss the question of whether the field possesses the right combination of social factors that make benchmarking a promising method at this point in time. The conclusions point to significant opportunities for collaborative benchmarks and systematic evidence sharing, but also several major challenges ahead.

General Terms

benchmark, digital preservation, software quality

Keywords

benchmark, digital preservation, software quality

1. INTRODUCTION

The number of different research results developing various preservation tools such as JHove2[1](characterization), Jpylyzer[2](quality assurance), Fido[3](identification) and others indicate their importance to the preservation community. The high quality of those tools is of major importance to the community. Although the community tends to acknowledge that better tools are still needed[4], proper evidence to support quality claims is still missing. This makes it hard to quantify the extent to which better tools are needed and how good the current ones actually are. Furthermore, the missing evidence puts major constraints on the decision making procedures which are implemented in various memory institutions.

Evidence, and the lack of it, has been a major concern in several fields closely related to digital preservation. Scientists have argued for experimentation, a type of empirical study, as an important method for providing evidence in software engineering and computer science [4][40]. However, different communities have shown different levels of acceptance of experimentation pointing to numerous reasons, such as costs and challenge to control all the variables, as a limiting barrier for rigorous adoption [40][26]. To address the barriers approaches such as testbeds and benchmarks have been proposed[26][3]. A benchmark is defined as “a standard against which measurements or comparisons can be made”[2]. A testbed is defined as “an environment containing the hardware, instrumentation, simulators, software tools, and other support elements needed to conduct a test”[2]. Even though both methods have comparison of software artefacts as their main goal slight difference can be distinguished. While a benchmark defines how the comparison should be done, a testbed is focused on providing a complete infrastructure to support that comparison. Tichy argued that benchmarks are an effective and affordable way to conduct experiments, although their development can require significant resources[40] .

In the digital preservation field, the term benchmark has been used several times but generally not accompanied by a rigorous treatment of the underlying assumptions, theories, requirements, limitations and techniques that are needed to make effective use of this method. This has resulted in several approaches which have not received sustained follow up. Benchmarks are thus still on the margin in the digital preservation field, even though this method has shown

[1]https://bitbucket.org/jhove2/main/wiki/Home

[2]http://jpylyzer.openpreservation.org/

[3]https://github.com/openpreserve/fido

[4]http://openpreservation.org/blog/2012/10/19/practitioners-have-spoken-we-need-better-characterisation/

major benefits in other fields.

The main contribution of this paper is the introduction of systematic and theory-based benchmarks to the digital preservation field. To enable the community to systematically define, use and evaluate benchmarks, a common model is required to define main benchmark components. Since the development of software tools is the focus, such a model should be based on theories from the software engineering field. Quality aspects of interest to our domain need to be backed up by well-defined quality models and metrics to enable objective comparison of the tools being benchmarked. Authenticity, as a key aspect of digital preservation, points to the correctness of tools as a crucial aspect of quality. However, this aspect has received insufficient effort so far[5]. Although the digital preservation community still lacks these benchmarks, several indicators signify the community's readiness.

This paper is organized as follows. In order to establish the basis for defining the common model, Section 2 provides an overview of the theory and practice of benchmarks in the software engineering and information retrieval fields. This is followed by an overview of related initiatives in the digital preservation field. Section 3 provides a common model for benchmarks. It defines the five main components of each benchmark. Section 4 provides four benchmarks which are described in terms of the five main components defined by the common model. Section 5 discusses the impact of proposed benchmarks and points to several preconditions which indicate community readiness for such benchmarks. Finally Section 6 summarizes the main conclusions and points to the future work.

2. THEORY AND PRACTICE OF BENCHMARKING

2.1 Benchmarks in related fields

The software engineering and information retrieval fields can be identified as most relevant fields for building benchmarks for digital preservation tools. One of the concerns of software engineering is to research and provide methods for evaluating software artefacts. Sim et al.[35]define a benchmark as "a test or set of tests used to compare the performance of alternative tools or techniques". Benchmarking has been a method employed by various laboratories and industries to objectively evaluate software solutions. The information retrieval field is mainly concerned with providing models and methods for an efficient information extraction from different sources. Digital preservation relies heavily on the metadata extracted from digital objects. This extraction, often performed by characterization tools, can also be considered to be a type of information retrieval.

Over the years research communities in software engineering and information retrieval have adopted and further developed benchmarks as a rigorous method to provide empirical evidence. This has provided an additional boost to the research and innovation in those fields. The Transaction Processing Council (TPC)[5] has been releasing a series of benchmarks covering various transaction actions. They have released over 750 benchmark publications covering a range of hardware and software platforms but have become most widely known for their database-centric benchmarks. The information retrieval field has several successful initiatives such as TREC[6], CLEF[7], MediaEval[8], and Mirex[9]. The Text Retrieval Conference (TREC), launched in 1992, has been releasing a number of information retrieval tasks organized in tracks to support evaluation of different information retrieval methodologies (in 2014 eight different tracks were organized). Numerous financial and nonfinancial benefits have been reported. It has been estimated that 16 million dollars of investment in TREC has resulted in 81 million dollars of extrapolated benefits[38]. The nonfinancial benefits are even more impressive ranging from providing large test collections and robust evaluation methodologies to enabling a competition which has fostered the whole research area. Many of the solutions have been adopted by the industry.

2.2 Components of a benchmark

In the software engineering field Sim et al.[35] propose a theory which views benchmarks as social and technical artefacts arising as the result of a consensus in a well-established community. Their interest is focused mainly on the technical research community. They have identified three major benchmark components: motivating comparison, task sample and performance measures, leaving open the order in which those components are developed.

- The **motivating comparison** defines the comparison to be done and the benefits that comparison will bring in terms of the future research agenda. For example, Kienle and Sim [19] motivate their benchmark for fact extraction from web sites by enabling the comparison of capabilities of different fact extractors. Heckman and Williams [16] propose a benchmark for tools that detect anomalies in source code. The main motivation is to find tools with the best rate of anomaly detection.
- The **task sample** is a list of tests that the subject, to which a benchmark is applied, is expected to solve. Kienle and Sim[19] use both artificial and real web sources as task samples for their web site extractors. Heckman and Williams[16] divide their task sample into two parts: six real Java subject programs and a list of true and false anomalies in those programs.
- The **performance measures** are qualitative or quantitative measurements taken by a human or a machine to calculate how fit the subject is for the task. For instance, Heckman and Williams[16] provide a list of well-established measures from the area of data mining and software anomaly detection.

In the information retrieval field Dekhtyar et al.[12] provide five main benchmark components: data set, tasks, answer set, measures and data representation formats/supplementary software. While tasks and measures are similar to the task samples and the performance measures proposed by Sim et al.[35], the typical usage scenario of information retrieval methods has identified data sets with accompany-

[5] http://www.tpc.org/information/about/abouttpc.asp

[6] http://trec.nist.gov/

[7] http://www.clef-initiative.eu/

[8] http://www.multimediaeval.org/

[9] http://www.music-ir.org/mirex/wiki/MIREX_HOME

ing answer sets as important benchmark components. The dataset contains information which a certain tool is required to retrieve. The answer set (often referred to as a ground truth) contains the correct answers which a tool is expected to return. It is reported by several authors that establishing a high-quality ground truth is the biggest challenge of such benchmarks[9][8] and the lack of it is a serious limiting factor [12]. Ben Charrada et al.[8] provide a real-world test dataset for which the ground truth is manually created. To reduce the impact of potential biases which could affect ground truth Chen et al.[9] proposed generating ground truth by a group of participants in several stages. Each stage is supposed to resolve conflicts from the previous stage.
The type of the data(tests) used in a benchmark plays an important role. Seng et al.[34] divide their database system benchmarks into two categories: synthetic and empirical. Synthetic benchmark create artificial data and tests, and empirical benchmarks use real world data and tests. They acknowledge that empirical benchmarks, even though ideal, in the case of databases are prevailed by synthetic due to the lower costs of implementing the synthetic ones.

To evaluate the benchmark quality several authors have proposed a list of desired characteristics. Sim et al.[35] propose a list of seven properties of successful benchmarks. Those are accessibility, affordability, clarity, relevance, solvability, portability and scalability. Huppler[17] proposes a list of five characteristics: relevant, repeatable, fair, verifiable and economical. He stresses repeatability as an important criterion allowing interested parties to get the same result even after repeating the whole benchmark. This criterion contributes to the overall trust in the results provided by a benchmark.

The provided components have shown to be beneficial in both fields as they allowed researchers to provide more focused benchmarks. It is clear that providing a common structure makes easier definition and comparison between similar benchmarks.

2.3 Awareness of the digital preservation community

The NDSA National Agenda for Digital Stewardship 2015 [27] highlights the importance of repeatable case studies and experiments, which are eventually to be transformed into “production public test beds” and “conformance tests”. The authors highlight that digital preservation is missing systematic metrics and measurements for “even simple failure scenarios”, which are dedicated to bit preservation.

To our knowledge, the first mention of the problem of lacking benchmarking in digital preservation is dated to 2000, when Greenstein[15] identifies benchmarking as an upcoming challenge for digital libraries. One of the early initiatives to create testbeds was carried out within the project Testbed Digitale Bewaring (Dutch Digital Preservation Testbed)[30] in 2002. The aim was to create testbeds for controlled experiments on preservation approaches (migration, emulation, XML) which were planned to be used by the Dutch government. As an example, the authors consider migration of MS Word documents within the testbed. They were interested to study documents features that change during the migration process. During the same time period, the development of testbeds was a key component of the US Digital Library Initiative (DLI) which led to the development of the D-Lib Test Suite[24].

The next milestone was the DELOS Digital Preservation Testbed, created in the DELOS project[37] in 2006. This testbed was based on the Dutch Digital Preservation Testbed. It contained a workflow of 14 steps, which were introduced to simplify the process of benchmarking, to guide users and to automate collection of evidence and documentation.

In 2007, Neumayer et al.[28] describe a range of issues arising when creating a testbed for digital preservation based on the accumulated experience and knowledge in the DELOS project. The challenges were (1) precise task definition, (2) definition of “sufficient” size of a benchmark, (3) benchmark samples generation, (4) data representation, and (5) ground truth and evaluation criteria specification. The authors attempted to empirically generalize on requirements and criteria, fleshing out a common structure of a benchmark.

Creation of the Planets Testbed[25] was inspired by the work undertaken by Dutch and DELOS testbeds in 2010. One of the critiques of the previous works was reliance on manual processes when characterising objects for a testbed. It is a time-consuming and error-prone activity, which is hardly applicable to large collections. The testbed here did not represent an actual real-world setting, but a software environment to explore with, test, and compare preservation tools and services in an online environment. These were open-ended tests, not necessarily focused on performance measures used for ranking tools. In parallel, the well-known decision support tool Plato for preservation planning process was developed[6]. In Plato, the focus is on systematic evaluation for the purpose of ranking and selection, and a strong emphasis is put on measuring and controlling the environment variables that influence results[7]. This makes the experiments rigorous, but the focus is situated on the particular decision making environment of one organization, and the requirements are tailored to these specific needs.

In 2011, the SCAPE project continued the work done on Plato in Planets, but adopted a different approach on the creation of the test environment. The project used its partners as sources for testbeds which were addressed by scenarios and constitute triplets of the following concepts: a dataset, a preservation issue and a possible solution[13]. This allowed them to structure the testbeds and think of potential scenarios and use cases, with limits on generalizability. Although the process of generation of datasets was automated, there is no confidence that the ground truth was valid and correct.

This issue is being addressed in the BenchmarkDP project. It is developing an approach to create benchmark datasets for objective validation of properties, such as functional correctness, of preservation tools [5]. Moreover, this approach allows automated generation of evidence for validity of datasets and corresponding ground truth.

2.4 Observations

As discussed earlier in this section, although there have been initiatives to address some specific cases for benchmarking, a holistic analysis of this challenge or at least an explicit list of benchmarks required in the digital preservation community does not yet exist. The work performed by the projects in digital preservation is lacking theoretical grounding (such as by Sim et al.[35]), so it is hard to rigorously evaluate requirements and criteria and study limitations of the testbeds.

Despite the existing efforts to create benchmarks and testbeds, there is still a deficiency of tests in digital preservation[33]. Hutchins[18] provided a thorough report on testing characterization tools. He confirms an issue of lacking ground truth datasets and methods, which would make it possible to verify correctness of a standalone tool. Rosenthal[31] also mentions lack of benchmarks in the bit preservation domain. He proposes strategies to improve competition in the market of software tools for bit preservation. The strategies are (1) agreement on common metrics, (2) consensus on modeling techniques for the metrics, (3) generation of better data and metadata, and (4) decreasing human factors as a reason for data loss. These strategies are applicable to the case of digital preservation as well: there is neither any agreement on metrics, nor ways to model these metrics, nor common approaches to create data for benchmarks.

These limitations prevent rigorous testing of the produced software tools. The community is aware of the shortcomings and define them as challenges in the research agenda. Practitioners are becoming aware of potential issues of selecting proper, trustworthy and correct components during decision making.

The benchmarking theory and practices from the other domains explained in this section are the foundations of the proposed approach to create benchmarks. The theory by Sim et al.[35] on benchmarks is a crucial pivot around which the body of benchmarks is to be built. It provides all necessary concepts and models which link the concepts and properties of successful benchmarks.

3. BENCHMARKS IN DIGITAL PRESERVATION

This section proposes a common benchmark model for digital preservation. The digital preservation tool benchmark defines a standardized way to objectively compare various software tools relevant to the digital preservation community. The common benchmark model defines five major components that each benchmark should define. As the focus of this paper is software tools, the model is not meant to be applicable to other areas of digital preservation where benchmarks might be used (e.g. organization benchmarks).

3.1 A common model for benchmarks in digital preservation

The theoretical work proposed by Sim et al[35] forms the basis for the common benchmark model. Based on the three proposed components (motivating comparison, task sample, performance measures) and the importance of data to the digital preservation community, five main benchmark components are identified: (1) motivating comparison, (2) function, (3) dataset, (4) ground truth (optional), and (5) performance measures.

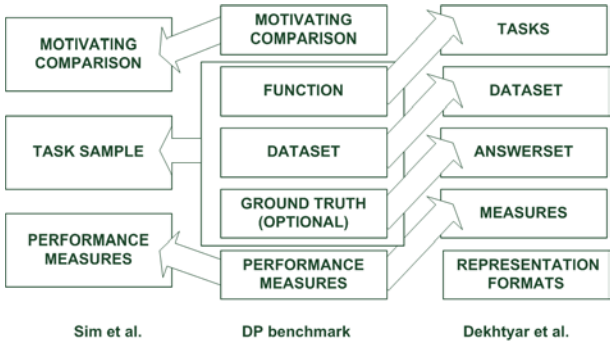


Figure 1: The common benchmark model mapped to the models from the SE and IR community

The *motivating comparison*, as defined by Sim et al. [35], will provide details on what a benchmark is supposed to compare. This can cover a variety of scenarios such as comparing tools in calculating significant properties values from electronic records, comparing different PDF validators or comparing different web harvesters in harvesting web pages. Each benchmark should motivate a comparison which is important to the community and is expected to further the whole research field.
The task sample proposed by Sim et al[35] has been divided into three parts: *function*, *dataset* and *ground truth*.
Function defines a specific task. It can range from migrating an object from one format to another to calculating values of a specific set of properties from a digital object.
The dataset defines a set of digital objects on which the specified task is to be executed. The dataset can be a set of images or documents, but also a set of software components (e.g. a set of video games which might be used in different emulation environments). To enable credible evaluation, in some cases the dataset might be accompanied by an appropriate ground truth.
The ground truth contains correct answers that a certain tool is expected to produce. For some motivating comparisons and task samples this element will not be required.
Performance measures demonstrate the fitness of the benchmarked tool for a certain task. As proposed by Sim et al[35], those measures can be quantitative or qualitative and can be calculated by a human or a machine. Performance measures are benchmark-specific which requires for each benchmark to properly document them together with the criteria for selecting them.

The common benchmark model can be unambiguously mapped to the models proposed in the software engineering[35] and information retrieval[12] fields (Figure 1).

3.2 What to compare and how to measure it? Quality modeling and performance measures

The main goal of the motivating comparison is to provide details on what a benchmark is supposed to compare. This can include various aspects such as the speed of a tool, usability or correctness of output. These quality aspects should be backed up by a quality model to avoid any misinterpretations and improve the clarity of a benchmark.

Table 1: A simple scenario mapped to the common model

Element	Question	Example
Motivating comparison	What to compare?	Correctness of characterization tools when extracting text from files.
Function	Which function?	Extraction of text from files.
Dataset	Which dataset?	MS Word files.
Ground truth	What is the ground truth?	Text inserted into each MS Word file.
Performance measures	What is calculated?	Percentage of files where text was correctly extracted.

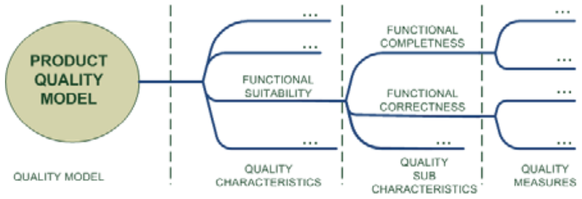


Figure 2: Hierarchical quality model

The ISO SQUARE Product Quality Model [1] organizes quality aspects such as speed, usability, and correctness into eight quality characteristics which are further divided into subcharacteristics. The software quality characteristics and subcharacteristics are indicated by one or more software quality measures[1].

Figure 2 shows the hierarchical decomposition of the Product Quality Model with the most relevant characteristics pointed out. Authenticity, the key concern of digital preservation, is considered when deciding on the relevance of characteristics. The concern is defined as "a degree to which a person or a system regards an object as what it is purported to be"[39]. Various tools capable of manipulating digital objects (e.g. migration) or measuring the values of object properties have the biggest impact on the authenticity of the digital object. Arguably the correctness of such tools is the most important quality aspect. The characteristic Functional Suitability and its two subcharacteristics Functional Completeness and Functional Correctness are identified as the most important characteristics related to authenticity. Those cover the degree to which a certain tool covers all the needed tasks and produces correct results[1].

The mentioned quality model provides a link between the motivating comparison and performance measures. The link is helpful to validate a selection of measures that are used to address a tool's quality. This will contribute to the clarity of benchmark specifications. An example of such linking is shown in Table 1 where the characterization tool's correctness is indicated as a percentage of files where the characterization task was successful.

As acknowledged by Sim et al. [35], creating performance measures (software quality measures) is particularly difficult. The digital preservation field has a systematic list of relevant quality measures [10] based on an ontology[23]. To expand this the information retrieval field with its numerous quality measures can be considered[36].

4. A SET OF BENCHMARKS

In this section, two benchmarks are introduced in details to demonstrate the applicability of the theory proposed previously. Additionally, there is a description of other benchmarks in Table 3. The proposed benchmarks are composed of the five components defined by the common model in Section 3.1. It is expected that each benchmark satisfies desired qualities defined by Sim et al[35]. However, due to the limited space the main focus of discussion is the relevance of the proposed benchmark to the digital preservation community and affordability. Thus the main goal of each benchmark definition is to provide a clear motivation to the digital preservation community and an understanding of what the benefits would be to the community when the benchmark is created and used. Furthermore, each benchmark will provide a clear and concise overview of the main tool function to be compared, requirements for the dataset, the nature and structure of the ground truth and an overview of applicable performance measures. Finally, each benchmark specification discusses major challenges that are expected when implementing the benchmark.

4.1 Raw photograph migration to DNG

4.1.1 Introduction

Raw photographs are images made by cameras and stored in a raw format. When considering digital preservation of raw photographs[22], migration is the most suitable candidate. This approach helps to avoid a risk of information loss due to discontinuation of support from a manufacturer. There are currently many proprietary raw formats with an undetermined lifetime. A common strategy to overcome this issue is migration to an open-source and standardized format. In this case, the format is DNG (Adobe Digital Negative). There are tools that allow such migrations like Photoshop[11], DNG Converter[12], CaptureOne[13], DigiKam[14] etc. Their application promises operational short-term benefits of homogeneous datasets that are easier to manage, as well as long-term benefits of lower risks of losing access to the assets. However, usually there is no evidence or confirmation based on rigorous testing that the tools work correctly during execution of a migration process. Therefore, the tools are not trustworthy and using them in preservation operations is risky. This benchmark enables the ranking of migration tools against a dataset of raw photographs. This is a practical problem for professionals and institutions, who consider selecting the best tool for raw photographs.

4.1.2 Motivating comparison

The purpose of this benchmark is a comparison of correctness of migration processes done by various software tools on the photograph dataset. It will show how similar a migrated is to the original photograph in terms of an image content,

[10] purl.org/dp/quality
[11] http://www.adobe.com/
[12] http://www.adobe.com/
[13] http://www.phaseone.com/
[14] http://www.digikam.org/

not metadata. Kulmukhametov et al.[22] discussed technical challenges when calculating similarity of photographs and introduced a tool, which will be used in this benchmark. The tool implements an algorithm which calculates Structure Similarity (SSIM) measure, which is claimed to be the closest measure to human perception when considering similarity of two images.

4.1.3 Function

The function to benchmark is migration of photographs from proprietary raw formats to the DNG format. As the files store raw data, tools usually do not provide many adjustable parameters, which would affect the content. Image compression is the only setting provided by the migration tools. As the goal of this benchmark is to test a correctness of migration, compression may significantly reduce the overall quality of the resulting photograph. This feature must be turned off.

4.1.4 Dataset

The dataset consists of photographs stored in raw formats. As the task is to compare correctness of migration tools, the dataset consists of photographs produced by different cameras and manufacturers. Such a dataset allows one to rank tools and determine the most versatile and universal one. Populating the dataset with photographs is achievable by using a content profiling tool C3PO[21], which allows one to extract samples from a bigger collection based on a specified list of criteria: a raw format, a camera model, a manufacturer.

4.1.5 Performance measures

The correctness of the tool is measured by calculating the SSIM value of a migrated photograph. The value is measured from 0 to 1. A higher magnitude of the value means better results. It is possible to compare the values from different tools for one photograph of the dataset. This makes it possible to identify the best tool for migration of this digital object. Another possibility is to calculate statistics based on the results of running the migration process for the whole dataset by one tool. The statistics, such as mean, median and standard deviation, may be helpful to identify the most versatile tool which produces the best results for the dataset.

4.1.6 Discussion

There is a challenge associated with this benchmark. It is about which photographs will constitute the dataset. There is no simple answer as the population of photographs is unknown. One possible solution is to provide samples of photographs created by different cameras from different manufacturers. Focusing on specific situations based on the requirements of the community is an important contribution to solve this challenge.

4.2 Property extraction from documents in electronic records environments

4.2.1 Introduction

Electronic records cover a spectrum of different use cases such as emails, audio or video records or documents. Document-based electronic records furthermore can cover a variety of scenarios such as books, articles or contracts. In many of those scenarios, document authenticity is of key importance. A migration tool can affect authenticity of a document by falsely migrating or not migrating at all some document elements. The lack of proper evidence around these cases makes it challenging to demonstrate authenticity of a document created by a migration.
To provide evidence for document authenticity, values of various document properties are measured. Pairs of property and value form a characteristic [11]. Stakeholders often point to significant properties of a document as important for its authenticity[29]. Expressing those properties in a measurable form enables assertion of document authenticity.

A number of different characterization tools, such as Apache Tika[15], National Library New Zealand Metadata Extractor[16] or Jhove2[17] claim to be capable of measuring values of various document properties. As they cover the commonly used formats such as MS Word and PDF they are suitable for providing evidence that is important for document authenticity. However, the coverage of needed properties and the correctness of measured values is not fully covered by a rigorous evaluation. This still hampers the validation of document authenticity as it is not possible to establish the confidence in the measured values.

Therefore, a benchmark is proposed to enable a rigorous evaluation of characterization tools when measuring document property values.

There are several major benefits of such a benchmark. The most important benefit is that it would provide the needed evidence around the quality of different characterization tools and enable an objective comparison of them. Furthermore, it is expected that it would foster the future development of those tools which would lead to better characterization tools. This would also be beneficial for establishing proper migration benchmarks which would be able to rigorously evaluate migration tools. As highlighted by Ross, ”before we can see migration as a viable aid to preservation, more work is needed in the development of metrics for benchmarking and supporting the evaluation of the risks or losses resulting from particular changes”[32].

4.2.2 Motivating comparison

The purpose of this benchmark is to enable the comparison of characterization tools with respect to the coverage of document properties and correctness of measured values for those properties. Coverage can be mapped to the functional completeness quality characteristic and correctness to the functional correctness. The Functional completeness is included mainly to denote if a certain tool can measure a property value. It is expected that in some cases some properties will not be fully covered which makes it an even more important aspect to systematically evaluate and compare.

4.2.3 Function

The main function is measuring values of document properties. Due to their importance for the authenticity, significant

[15] http://tika.apache.org/
[16] http://meta-extractor.sourceforge.net/
[17] https://bitbucket.org/jhove2/main/wiki/Home

Table 2: Quality characteristics and performance measures

Quality Characteristic	Measure	Calculated as
Functional completeness	Coverage	calculated per property as a percentage of documents where a tool returned a value for a specific property
Functional correctness	Accuracy	calculated per property as a percentage of files where a tool returned correct value for a specific property
	Exact Match Ratio	calculated for a tool as a percentage of files where the tool returned correct values for all properties

properties are in the focus of the benchmark. However as pointed out by Dappert et al.[11] the significance of a property is not absolute and binary but depends on the stakeholders' requirements for a certain document(or a collection of documents). Thus it will be challenging or even impossible to come up with a list of required significant properties. However, it can be argued, due to the similar scenarios various content holders are dealing with, that it is possible to come up with a list of commonly used properties which are identified as significant for documents in electronic records environments. Building on previous studies that classified and modelled significant properties in preservation planning case studies, and by analyzing actual preservation plans created by different stakeholders a list of common significant properties can be made.

4.2.4 Dataset and ground truth

In order to cover different documents types the dataset should be focused on the combinations of different document elements and their properties. Here, document elements denote simple building blocks which are used to compose a document (pages, footers, text, images) and their properties such as font color, table size, and image position. This affects the size of the needed dataset. The bigger the number of elements and their properties covered, the bigger the dataset. Even in a very simplistic scenario with five elements where each element has three properties with only one possible value the dataset would need to contain 125 documents to cover all the combinations. The real world is much more complex with more elements, properties and their possible values. This combinatorial explosion makes automatic dataset generation a better method, than the manual annotation, for establishing a proper dataset.

To enhance automation the ground truth needs to be expressed in a machine-readable form. It should specify the correct property-value pairs.

4.2.5 Performance measures

This benchmark addresses two quality characteristics (the functional completeness and the functional correctness). Each characteristic is indicated by one or more measures.

The functional completeness is covered by one measure. This measure should point out how well a single tool covers defined properties. Therefore for each property a percentage is calculated to show the number of documents where a tool returned a value for specific property.

When dealing with functional correctness, there are two aspects that are important to consider. There is the need for a measure that will show how good a tool is on the whole set of properties and on a specific property. For example, it can be important to know that a specific tool which does not have good overall performance has remarkable performance on one of the properties.

4.2.6 Discussion

The proposed benchmark would bring several benefits to the digital preservation community. It would enable an objective comparison of characterization tools in terms of their coverage and correctness when measuring significant properties from documents in electronic records environments. This would provide objective evidence and drive the future development of tools.

The biggest challenge of this benchmark is the dataset generation. Its combinatorial growth, dependent on the number of elements and properties, makes manual annotation insufficient as a method for dataset generation. Automatic dataset generation should provide efficient methods to model different documents in terms of their possible elements and how to control the combinations of those elements and their properties. The model-driven engineering framework[5] provides a possible solution for this problem. The feasibility of the approach has been demonstrated on a similar scenario. However, future work will be required to enhance the whole method to be more robust and cover a larger number of elements.

Once created, it is expected that the effort required for running the benchmark will not be significant. The dataset, even though expected to have a significant number of objects, is still expected to be in the range which standard commodity hardware can handle. Using an artificial dataset raises some issues around the relevance of the benchmark. The biggest challenge that the generation method will need to address is the representativeness of the generated dataset of real-world datasets.

4.3 PDF validation and Web harvesting benchmark

Due to limited space, two additional benchmarks are presented in Table 3. The two benchmarks cover the scenarios of PDF validation and web harvesting.

The PDF file format family has been proliferated over the years as the defacto standard for storing and exchanging various kinds of documents(articles, books, ...). The quality of available validators, used to check the validity of a PDF file, is diverse and hard to objectively compare. Initiatives to build even more validators[18] show that the community is still not satisfied with existing offerings. This points to

[18] http://openpreservation.org/news/verapdfa-consortium-awarded-phase-1-of-preforma-call-for-tender-for-pdfa-validation/

Table 3: PDF validation and Web harvesting benchmarks

Name	PDF validation	Web harvesting
Motivating comparison	Compare validation functional correctness of different PDF file format validators. Furthermore compare the functional correctness of reported violations	Compare functional correctness and completeness of a web harvester
Function	Validate a PDF file	Harvest a web site
Dataset	PDF files covering valid and invalid examples. Invalid examples cover various combinations of violations	A set of webpages. Web-pages are accessed by providing a GET request to a web-server. The settings of the server are set in the benchmark.
Ground truth	Information pointing to the true validity of a PDF file. In the case of an invalid file provides the true violations expected to be reported from a tool	A list of properties for each web-page in the data set: size of the web-page, HTTP GET request, html markup, presence of resources and executable scripts
Performance measures	Accuracy of a validation output; Accuracy of reported violations	Correctness and completeness of the web harvesting tools are measured by calculating precision for the properties

the need for a proper benchmark to enable a proper tool evaluation and comparison. The benchmark would provide an objective evaluation of PDF validation tools.

Web harvesting is an important function in the web archiving community. However due to the complexities of current web pages in terms of links and various technologies being used (e.g. JavaScript and Flash) it is hard to understand the completeness and correctness of the harvesting task.
The proposed benchmark should therefore enable rigorous testing of web harvesting tools by focusing on aspects such as the use of JavaScript, Flash or complex linking structures (spider traps).

5. DISCUSSION

5.1 Preconditions and success factors

Benchmarking as a rigorous method is not a simple, easily completed task. Does our community meet the required preconditions for benchmarking? It is worth revising the requirements and success factors highlighted by Sim [35].
Benchmarks should be collaborative, open, and public. The community has a long track record of sharing various forms of knowledge; however, this has not been replicated when it comes to sharing data. Despite efforts such as LDS3[19], the OPD data endpoint[20], and isolated data sets such as from the UK Web Archive[21], data sharing is not common for a number of reasons. We hope to address some of this by generating data that can be shared freely.

The community must be ready to incur the costs of benchmarking. Continued evolution of the benchmark will be necessary. It will require a selective approach with a focus on those motivating comparisons that are truly encapsulating the paradigms of the field to catalyze substantial interest of the community.

Benchmarks encapsulate paradigms. Benchmarks must be developed by consensus. Are our paradigms understood well enough?

Design decisions need to be supported by lab work. Benchmarking needs to use established results where possible. We base the existing work and proposals in this paper on extensive lab work and case studies in preservation planning and beyond.

Choosing the task sample may be controversial. Consensus is needed in the community, and efforts as part of BenchmarkDP are focused on outreach and community engagement.

The community must have an opportunity to participate, provide feedback, and endorse benchmarks. Efforts should be led by a small number of champions. IPRES as the leading conference in the field is the ideal place for engagement and participation. The authors encourage interested community members to get involved.

5.2 Datasets and ground truth: the key challenge

There is a general lack of test datasets with acompanying ground truth for preservation tools. The widely known and used dataset is the Govodcs dataset[22][14]. However, the only available ground truth is related to identification data. Since that data has been produced by a forensics tool, provided by Forensics Innovations[23] the validity of the ground truth is hard to confirm. Furthermore, the whole dataset is applicable to a limited range of identification scenarios.
Two main approaches in creating test datasets are identified: 1) *subsampling real world datasets and manually annotating them*, and 2) *automatically generating datasets with an accompanying ground truth*.

Manual datasets annotation brings one obvious advantage. Using a real-world test sample makes the benchmark relevant to the real-world scenarios and as such the benchmark results are more trustworthy. However, producing such datasets will be an effort-intensive job and datasets will need to be reduced to a smaller number of objects to make manual annotation plausible. In order to remove any kind of unwanted biases, automatic methodologies for analysing and subsampling real datasets are required. The content

[19] `http://beta.lds3.org/`

[20] `http://wiki.opf-labs.org/display/PT/The+OPF+Data+Endpoint`

[21] `http://data.webarchive.org.uk/opendata/ukwa.ds.2/`

[22] http://digitalcorpora.org/corpora/files

[23] http://www.forensicinnovations.com/

profiling tool C3PO[24] provides a scalable architecture for automatic content profiling of digital objects. It thus, provides the basis on top which sampling algorithms can be built [21][10]. For some functions this kind of sampling from a real world dataset will be sufficient and was the common approach until now in the community.

In some cases e.g. where detailed annotations about technical details are required or the number of features or their combinations require significant size of a sample set and manual annotation might still be too expensive or even impossible. In those situations automatic datasets generation is a possible approach. While in other fields this approach is already researched, the digital preservation field has only started to explore its possibilities and the approach is considered to be highly novel[5][20]. Becker and Duretec [5] proposed a framework based on Model Driven Engineering principles for automatic test dataset generation. This framework has been the basis for several prototypes that serve as a proof of concept. However this is a novel approach in the digital preservation and as such will require significant research effort.

6. CONCLUSION AND FUTURE WORK

Much of the research effort in digital preservation is invested in developing software tools for managing, processing and disseminating digital information. The community has increasingly recognized the need for systematic testing and evidence sharing on different characteristics of quality of those tools. In this article, we introduced insights from theory and practice of benchmarking of Software Engineering and Information Retrieval communities and discussed how the introduction of systematic benchmarking provided a boost for research and innovation in these communities. Based on a simple framework for specifying and analyzing benchmarks, we outlined a set of initial specifications for benchmarks. While this initial set is by no means complete, it provides a key stepping stone towards collaborative campaigns for benchmarking. The defined four benchmarks will be a starting point for community involvement in establishing benchmarking in digital preservation as an important method for strengthening the evidence base.

An essential characteristic of a successful benchmark is that it will lead to better tools, to the point that a majority of tools complete standard benchmarks with near-perfect scores. This means that it is possible to start with quick wins for comparison tasks that are comparably simple, but relevant for comparison, roadmap generation and prioritization of future development of tools, in order to establish the mechanisms of benchmarking as a method; and then proceed to advanced, more challenging benchmarks as experience accumulates.

But more importantly, it means that each successful benchmark will eventually be superseded by an evolved specification. It will require joint community interest and efforts to make such efforts feasible and worthwhile; and hence, a focus is needed on those quintessential tasks for which a systematic, rigorous comparison of candidate components on a widely agreed performance measure is possible, necessary, and relevant. It is up to the members of the community to ensure that their needs are part of this consensus.

[24] http://ifs.tuwien.ac.at/imp/c3po

Acknowledgements

Part of this work was supported by the Vienna Science and Technology Fund (WWTF) through the project *BenchmarkDP* (ICT12-046).

7. REFERENCES

[1] *ISO/IEC 25010 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models*. 2010.

[2] Systems and software engineering – Vocabulary. *ISO/IEC/IEEE 24765:2010(E)*, pages 1–418, Dec. 2010.

[3] E. Barreiros, A. Almeida, J. Saraiva, and S. Soares. A Systematic Mapping Study on Software Engineering Testbeds. In *2011 International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 107–116, Sept. 2011.

[4] V. Basili. The role of experimentation in software engineering: past, current, and future. In *, Proceedings of the 18th International Conference on Software Engineering, 1996*, pages 442–449, Mar. 1996.

[5] C. Becker and K. Duretec. Free Benchmark Corpora for Preservation Experiments: Using Model-driven Engineering to Generate Data Sets. In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, pages 349–358, New York, NY, USA, 2013. ACM.

[6] C. Becker, H. Kulovits, A. Rauber, and H. Hofman. Plato: A service oriented decision support system for preservation planning. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 367–370. ACM, 2008.

[7] C. Becker and A. Rauber. Improving component selection and monitoring with controlled experimentation and automated measurements. *Information and Software Technology*, 52(6):641–655, 2010.

[8] E. Ben Charrada, D. Caspar, C. Jeanneret, and M. Glinz. Towards a Benchmark for Traceability. In *Proceedings of the 12th International Workshop on Principles of Software Evolution and the 7th Annual ERCIM Workshop on Software Evolution*, IWPSE-EVOL '11, pages 21–30, New York, NY, USA, 2011. ACM.

[9] X. Chen, J. Hosking, J. Grundy, and R. Amor. Development of Robust Traceability Benchmarks. In *Software Engineering Conference (ASWEC), 2013 22nd Australian*, pages 145–154, June 2013.

[10] Christoph Becker, Luis Faria, and Kresimir Duretec. Scalable decision support for digital preservation. *OCLC Systems & Services: International digital library perspectives*, 30(4):249–284, Nov. 2014.

[11] A. Dappert and A. Farquhar. Significance is in the Eye of the Stakeholder. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL'09, pages 297–308, Berlin, Heidelberg, 2009. Springer-Verlag.

[12] A. Dekhtyar and J. Hayes. Good Benchmarks are Hard To Find: Toward the Benchmark for Information Retrieval Applications in Software Engineering. *Information Retrieval in Software Engineering, International Conference on Software Maintenance (ICSM): Philadelphia, PA.*, Sept. 2006.

[13] M. Ferreira, H. Silva, R. Castro, P. Moldrup-Dalum, Z. Pehlivan, C. Wilson, and S. Schlarb. D10.2 gap analysis on action services tools and scape platform and testbeds requirements, 2013.

[14] S. Garfinkel, P. Farrell, V. Roussev, and G. Dinolt. Bringing Science to Digital Forensics with Standardized Forensic Corpora. *Digital Investigation*, 6:S2–S11, Sept. 2009.

[15] D. Greenstein. Digital libraries and their challenges. *Library trends*, 49(2):290–303, 2000.

[16] S. Heckman and L. Williams. On Establishing a Benchmark for Evaluating Static Analysis Alert Prioritization and Classification Techniques. In *Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '08, pages 41–50, New York, NY, USA, 2008. ACM.

[17] K. Huppler. The Art of Building a Good Benchmark. In R. Nambiar and M. Poess, editors, *Performance Evaluation and Benchmarking*, number 5895 in Lecture Notes in Computer Science, pages 18–30. Springer Berlin Heidelberg, 2009.

[18] M. Hutchins. Testing software tools of potential interest for digital preservation activities at the national library of australia. *National Library of Australia Staff Papers*, 2012.

[19] H. Kienle and S. Sim. Towards a benchmark for Web site extractors: a call for community participation. In *Seventh European Conference on Software Maintenance and Reengineering, 2003. Proceedings*, pages 82–87, Mar. 2003.

[20] Y. Kim and S. Ross. Searching for Ground Truth: A Stepping Stone in Automating Genre Classification. In C. Thanos, F. Borri, and L. Candela, editors, *Digital Libraries: Research and Development*, number 4877 in Lecture Notes in Computer Science, pages 248–261. Springer Berlin Heidelberg, 2007.

[21] A. Kulmukhametov and C. Becker. Content Profiling for Preservation: Improving Scale, Depth and Quality. In K. Tuamsuk, A. Jatowt, and E. Rasmussen, editors, *The Emergence of Digital Libraries – Research and Practices*, number 8839 in Lecture Notes in Computer Science, pages 1–11. Springer International Publishing, Nov. 2014.

[22] A. Kulmukhametov, M. Plangg, and C. Becker. Automated quality assurance for migration of born-digital images. In *Archiving Conference*, volume 2014, pages 73–78. Society for Imaging Science and Technology, 2014.

[23] H. Kulovits, M. Kraxner, M. Plangg, C. Becker, and S. Bechhofer. Open preservation data: Controlled vocabularies and ontologies for preservation ecosystems. *Proc. IPRES*, pages 63–72, 2013.

[24] R. L. Larsen. The dlib test suite and metrics working group: Harvesting the experience from the digital library initiative. *D-Lib Working Group on Digital Library Metrics Website*, 2002.

[25] A. Lindley, A. N. Jackson, and B. Aitken. A collaborative research environment for digital preservation-the planets testbed. In *Enabling Technologies: Infrastructures for Collaborative Enterprises (WETICE), 2010 19th IEEE International Workshop on*, pages 197–202. IEEE, 2010.

[26] M. Lindvall, I. Rus, F. Shull, M. Zelkowitz, P. Donzelli, A. Memon, V. Basili, P. Costa, R. Tvedt, L. Hochstein, S. Asgari, C. Ackermann, and D. Pech. An evolutionary testbed for software technology evaluation. *Innovations in Systems and Software Engineering*, 1(1):3–11, Mar. 2005.

[27] National Digital Stewardship Alliance. 2015 National Agenda for Digital Stewardship, 2015.

[28] R. Neumayer, H. Kulovits, M. Thaller, E. Nicchiarelli, M. Day, H. Hofmann, and S. Ross. On the need for benchmark corpora in digital preservation. In *Proceedings of the 2nd DELOS Conference on Digital Libraries*, 2007.

[29] Parliamentary Archives. *A Digital Preservation Policy for Parliament*. London, Parliamentary Archives, 2009.

[30] M. Potter. Researching long term digital preservation approaches in the dutch digital preservation testbed (testbed digitale bewaring). *RLG DigiNews*, 6(3), 2002.

[31] D. S. Rosenthal. Bit preservation: A solved problem? *International Journal of Digital Curation*, 5(1):134–148, 2010.

[32] S. Ross. Changing Trains at Wigan: Digital Preservation and the Future of Scholarship, Jan. 2000.

[33] R. Ruusalepp and M. Dobreva. Digital preservation services: State of the art analysis. 2012.

[34] J.-L. Seng, S. B. Yao, and A. R. Hevner. Requirements-driven database systems benchmark method. *Decision Support Systems*, 38(4):629–648, Jan. 2005.

[35] S. E. Sim, S. Easterbrook, and R. C. Holt. Using Benchmarking to Advance Research: A Challenge to Software Engineering. In *Proceedings of the 25th International Conference on Software Engineering*, ICSE '03, pages 74–83, Washington, DC, USA, 2003. IEEE Computer Society.

[36] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, July 2009.

[37] S. Strodl, A. Rauber, C. Rauch, H. Hofman, F. Debole, and G. Amato. *The DELOS testbed for choosing a digital preservation strategy*. Springer, 2006.

[38] G. Tassey, B. R. Rowe, D. W. Wood, A. N. Link, and D. A. Simoni. Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program. *National Institute of Standards and Technology, Gaithersburg, Maryland*, 2010.

[39] The Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*. June 2012.

[40] W. Tichy. Should computer scientists experiment more? *Computer*, 31(5):32–40, May 1998.

Deduplicating Bibliotheca Alexandrina's Web Archive

Youssef Eldakar
Bibliotheca Alexandrina
PO Box 138
Alexandria 21526
EGYPT
youssef.eldakar@bibalex.org

Magdy Nagi
Bibliotheca Alexandrina
PO Box 138
Alexandria 21526
EGYPT
magdy.nagi@bibalex.org

ABSTRACT

Archiving web content is bound to produce datasets with duplication, either across time or across location. The Bibliotheca Alexandrina (BA) has a web archive legacy spanning a period of 10 years and is continuing to expand the collection. Initial assessment of this very large store of data was conducted. Given a high enough rate of duplication, deduplication would lead to sizable savings in storage requirements. The BA worked through the International Internet Preservation Consortium (IIPC) to compile best practices for recording duplicates in ISO 28500, the WARC File Format. To deduplicate legacy web archives "after the fact," the BA is implementing the WARCrefs deduplication tools. Following implementation and testing, the BA plans to put the tools to use to deduplicate its one petabyte of archived web content.

General Terms

case studies and best practice

Keywords

web archiving, deduplication, hash algorithms, ISO 28500, WARC File Format, WARCrefs, WARCsum

1. INTRODUCTION

The International Internet Preservation Consortium (IIPC) defines *web archiving* as "the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use" [9]. During the collection phase of the process, a *crawler* is used to explore a network of hyperlinks, starting off at a set of seeds, fetching resources it visits. This process is typically repeated periodically to capture changes over time.

However, even though the web is quite a dynamic place, a resource will not necessarily be modified during the interval between one visit and a subsequent revisit. In addition, even though the web is quite a diverse place, a resource will not necessarily be unique within the web archive when compared to other resources at different locations. For the former situation, consider, for instance, the text of the constitution of some country on the government's website, which can be expected to remain unmodified for years, where archiving subsequent identical snapshots of the resource as-is introduces duplication across time into the archive. For the latter situation, consider, for instance, a photo of some event that is posted to a blog, a social networking site, as well as a personal homepage, where archiving all identical instances of the resource as-is introduces duplication across location into the archive.

The Bibliotheca Alexandrina (BA) in Alexandria, Egypt, has in its holdings a legacy web archive of broad web crawls provided by the Internet Archive in San Francisco. This archive of a decade's web history starting in 1996 plus BA's own collection of focused web crawls started in 2011 total approximately 80 billion records and are stored in approximately one petabyte in compressed form. For backup, an additional petabyte is required. And even though the data is hosted on a commodity hardware computer cluster, at such large scale, reduction in storage requirements even by a relatively modest percentage would lead to sizable financial savings and offer more room for expanding collection activities.

Beyond using storage more efficiently, the desire to deduplicate web archive data is driven by a few extra motivations. Kristinn Sigurdhsson, in a 2006 publication [16], takes a look at arguments for deduplication in a web archive. It is noted therein that reduction in storage requirements is the most notable benefit in addition to improving the quality of the collection or its presentation. Other benefits also mentioned therein but applying only to deduplication during crawl time based on HTTP headers are reduction of load on web servers as well as reduction in bandwidth consumption. We may also add to the benefits of deduplication improvement in performance on the access interface, as knowing which resources are duplicates of which resources would improve the caching implementation.

Could the rate of duplication in the BA web archive be significant enough that the benefits to be achieved merit the effort?

2. IDENTIFYING DUPLICATES

Before evaluating how much more efficient use of the storage infrastructure would become should the BA web archive be free of duplicates, a method for telling whether two resources are identical is needed. The most rudimentary one is to compare the data streams byte-by-byte. Given a set of n data streams, $n(n-1)/2$ comparisons will be required, because each data stream will have to be compared with all data streams in the set but itself, forming a complete graph with n vertices. To optimize, comparisons where the data streams are not equal in size and therefore cannot be identical will be skipped. To further optimize, comparisons will also be skipped where the data streams have already been found identical via an indirect route on the graph. Yet, where n is very large, and where the data is hosted on a distributed storage infrastructure, this method will not scale well, because data will have to be marshalled heavily on the network during the repetitive reads and compares.

An alternative method is *hashing*. Hash functions are algorithms that map a data stream of arbitrary length to a fixed-length *hash value*, which uniquely identifies the data stream [3]. Using this method, each data stream will be read once and hashed. Each hash value along with a reference to the data stream will be inserted into a list. The list is sorted on the hash value, clustering entries for identical data streams together. Costly repetitive reads and compares in the former method are replaced with a much lighter merge-sort.

Password verification, data integrity checking, and even automatic *deduplication* built into modern file systems such as ZFS [7] and Btrfs [1] are examples of today's common applications of hash algorithms. Of hash algorithms, MD5, SHA-1, and SHA-2 seem to be de-facto standards in the industry.

A hash algorithm is reliable up until it is shown to entail a risk for *collisions*, where two unidentical data streams are mapped to the same hash value. As presented in CWI's "Cryptanalysis of MD5 and SHA-1" [17], the possibility for collisions is demonstrated for MD5 in theory as well as in practice, and only in theory for SHA-1. To date, there seems to be no published work demonstrating susceptibility of SHA-2 to collisions. However, the more reliable SHA-1 and SHA-2 algorithms come at a cost. In Crypto++ 5.6.0 benchmarks [2], MD5 performed 65 percent faster than SHA-1, and well over twice as fast as 256-bit SHA-2 (SHA-256) and 512-bit SHA-2 (SHA-512). See Figure 1.

Falsely identifying resources in the web archive as duplicates shall not be tolerated, as that would lead to corruption in web history. Here, one model to take as reference is how deduplication is managed in the ZFS file system, where false positives also are not tolerated. ZFS uses SHA-256 to identify duplicates but also runs positives through collision resolution, which compares the files byte-by-byte to rule out collisions. Such a model seems quite suitable for application in web archive deduplication. In fact, in the safety of collision resolution, even MD5 could be considered for its significant speed advantage, and in hopes collisions will not be very frequent. Further, statistics on collisions in such very large dataset as a web archive could be of value to the cryptography community.

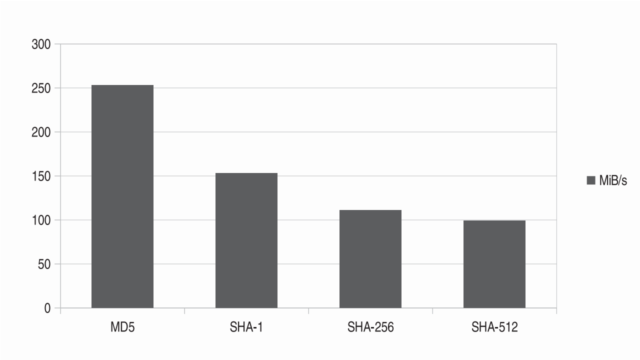


Figure 1: Crypto++ hash algorithm benchmarks.

3. INITIAL ASSESSMENT

In 2012, the BA sampled close to 10 percent of data in its web archive. Hash values for the resources were computed and sorted into a list. Out of this sample, the rate of duplication was found to be approximately 14 percent. With a bigger sample, it may be reasonable to optimistically hope the rate of duplication will be higher. Still, even just a saving of 14 percent within the petabyte of data in the BA web archive would translate into 140 terabytes of space. Multiplying this number by two to account for the backup, the total storage saved becomes 280 terabytes. Such saving on storage would yield a considerable cut-down on expenses, which may possibly be invested towards widening the scope of web archiving at the institution.

In addition to the duplication assessment on the BA web archive, other web archives also report significant rates of duplication. For instance, in a presentation during the 2011 IIPC General Assembly [15], the National Diet Library, Japan, estimated deduplication would reduce the Japanese monthly web archives by 80 percent, and the quarterly archives by 45 percent. Further experiences shared during discussions at IIPC meetings did also describe the rate of duplication as being significant in web archive collections at other institutions.

4. STANDARDIZING THE RECORDING OF DUPLICATES

The de-facto standard format used to store resources in web archives is the ARC File Format initially conceived at Alexa Internet [11], to which ISO 28500, the WARC File Format, is a more comprehensive successor [14].

In Sigurdhsson's 2006 publication [16], it is noted in the conclusion, "While there are difficulties in presenting collections that have been [deduplicated], the introduction of the WARC File Format should greatly alleviate that" [16]. However, even though the ISO specification provides for `revisit` records, the specifics for practical usage are not clearly outlined. In 2013, the BA worked through the IIPC Harvesting Working Group (HWG) to draft a "Proposal for Standardizing the Recording of Arbitrary Duplicates in WARC Files" [10]. Later that same year, the proposal earned IIPC Steering Committee adoption as recommended best practices.

The proposal recommends the use of the following fields in `revisit` records with the `identical-payload-digest` profile in the WARC File Format to replace duplicate resources with references to the initial capture:

`WARC-Refers-To-Target-URI`: The URI of the original resource.

`WARC-Refers-To-Date`: The date of the original resource.

In addition, the proposal discourages the use of "fields specifying the actual WARC file name and offset," as such usage is "potentially very brittle."

5. TOOLS

Deduplication of web archive data may be carried out at either of two phases: during a crawl, or after the crawl.

5.1 During a Crawl

In Sigurdhsson's 2006 publication [16], modules implemented for the Heritrix crawler [4] to stop processing of duplicate resources are presented. The `DeDuplicator` module depends on hash values to identify duplicates using an index of previously crawled resources as reference. In addition, the implementation also provides an alternative method for identifying duplicates based on the datetime and/or ETag in HTTP headers fetched during the crawl: the `DeDupFetchHTTP` module.

5.2 Post-Crawl

Even though the `DeDuplicator` and `DeDupFetchHTTP` modules are quite effective in eliminating duplicates during a crawl, the BA requires a different type of solution that enables the archive keeper to sort out duplicates "after the fact" in a legacy collection. For this, the BA has implemented the WARCrefs set of tools for identifying duplicates and converting them to references in a web archive collection after crawl time.

BA's developed solution is divided into two separate packages: WARCrefs for doing the deduplication, and WARCsum for generating hash manifests of web archive resources, doing collision resolution, and post-processing the manifests, which serve as input to WARCrefs. WARCrefs is implemented in Java, because well-maintained WARC manipulation APIs are available in this language. WARCsum, on the other hand, is implemented in C, because hash generation as well as collision resolution are time-consuming tasks, particularly when dealing with big data, which makes C as a lower-level language a good choice when seeking to tune performance. The software is operated at the command line.

Stage 1 in the deduplication process is `warcsum`. This tool takes as input a list of WARC files. For each record of type `response` in each file in the input, one line of output is written to the hash manifest. Each hash manifest line consists of six fields: the WARC file name, the file offset in bytes at which the record is located, the length of the record also in bytes, the URI the record captures, the date of the capture, and the hash value of the content in the payload, excluding headers. The following is an example of what `warcsum` writes for a record in the input (for readability, each field is on a separate line):

```
TGvr4fAfmc.warc.gz
3901
635
http://www.akhbarway.com/robots.txt
2012-04-08T20:13:38Z
sha1:aa20238aab9cea0696a9b5d5f7a44a42de16adfc
```

`warcsum` can be configured using command-line options. `warcsum` uses hash functions from OpenSSL [5]. MD5, SHA-1, SHA-256, and SHA-512 are supported. For records where a hash value is already present in the record headers, `warcsum` can be set to use the existing value or recompute the hash. Records with empty content can be skipped or treated as normal records.

`warcsum` is to be run on each host in the computer cluster that makes up the data store where WARC files are kept. Output from all instances is to be aggregated and sorted on the hash value field.

Stage 2 is `warccollres`, provided in the WARCsum package. This tool performs the collision resolution, acting as a safety measure against false positives due to hash collisions. `warccollres` takes as input the aggregated hash manifest generated by `warcsum`. For each cluster of lines having the same hash value, the content is fetched from the data store and compared byte-for-byte to verify whether the records are indeed duplicates. The result of the collision resolution is saved to the hash manifest line by appending a seventh field: a *hash extension*, which is a sequential number that distinguishes unidentical records incorrectly given the same hash value. The output from `warccollres` is an extended hash manifest.

To access the data store, `warccollres` looks up file names in a MySQL database to determine where they are located then fetches the records via HTTP. The de-facto standard web archive access system, the OpenWayback [6], already depends on HTTP for fetching records, hence going with HTTP as the first choice for the fetch method and reusing an existing infrastructure was natural. In the future, alternative fetch methods may be implemented into `warccollres`.

`warccollres` is to be run on one or more hosts with HTTP access to the data store. When run on multiple hosts to distribute the collision resolution workload, the input hash manifest is to be partitioned across each host such that each cluster of lines having matching hash values is fully contained within a single partition. `warcsumsplit` is a tool provided in the WARCsum package for this purpose. Output from all `warccollres` instances is to be aggregated and sorted on the hash value and hash extension fields.

Stage 3 is `warcsumproc`, also provided in the WARCsum package. This tool post-processes the extended hash manifest, further extending it such that each line encapsulates all the information the deduplication stage needs to operate on the record the line is for. Post-processing looks at the hash value and hash extension in each line in the context of the line before it and writes a *copy number* as the eighth field on the line. Thus, a line where the copy number is 1 is an original record to be kept intact, while a line where the copy number is greater than 1 is a duplicate to be converted into a reference, i.e., a **revisit** record. In addition, where the copy number is greater than 1, the post-processed hash manifest also has as the ninth and tenth fields the URI and date, respectively, of the original record, which is information needed to construct the **revisit** record. The post-processed hash manifest is to be sorted on the file name and offset fields.

The post-processing functionality is also available in **warccollres** and can be enabled using a command-line option, in which case, **warccollres** outputs a post-processed hash manifest. This is more efficient than performing the post-processing in a separate stage. However, given enough confidence that the hash algorithm being employed is not likely to have collisions, with some risk, the collision resolution may be skipped in order to save time, in which case, **warcsumproc** is needed. Needless to say, opting not to perform the collision resolution is quite inadvisable.

Both **warcsum** and **warccollres** read WARC files using **libgzmulti**, a library the BA developed as a wrapper around **zlib** [13] for working with multi-member GZIP files, of which WARC files are a type.

Stage 4 is where the post-processed hash manifest produced by the toolchain provided in the WARCsum package is turned over to WARCrefs to perform the deduplication. Similar to **warcsum** (stage 1), the **warcrefs** tool takes as input a list of WARC files but also now has access to post-processed hash manifest lines for records in the files it is to operate on. **warcrefs** iterates through each WARC file in the input and also concurrently through corresponding lines in the post-processed hash manifest. Each record with a copy number greater than 1 in the corresponding manifest line is converted into a **revisit** record, where **WARC-Refers-To-Target-URI** and **WARC-Refers-To-Date** in the record headers are set to the URI and date, respectively, of the original resource, and payload headers are transferred as-is into the **revisit** record. Otherwise, if the copy number is 1, or if no corresponding line is in the manifest, the record is not altered.

warcrefs uses the Java Web Archive Toolkit (JWAT) [12] for WARC file IO. **warcrefs** can be configured to rewrite files in-place or save to a new file.

warcrefs is to be run on all hosts in the data store. The post-processed hash manifest is to be split across the hosts such that each host only has lines corresponding to records in WARC files on the host. Further, as the absence of a manifest line for a record implies the record is not a duplicate, lines where the copy number is 1 are to be omitted to reduce the amount of manifest data **warcrefs** has to process.

Figure 2 illustrates the deduplication process.

WARCrefs, WARCsum, and **libgzmulti** will be available open-source on GitHub [8].

6. EXECUTION

With the solution implemented, the next objective is to put the WARCrefs deduplication tools to use to deduplicate the full BA web archive. The plan is as follows:

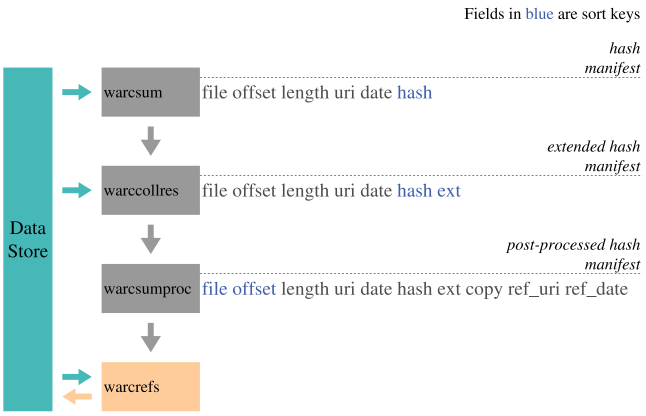

Fields in blue are sort keys
hash manifest
warcsum
file offset length uri date hash
extended hash manifest
warccollres
file offset length uri date hash ext
Data Store
post-processed hash manifest
warcsumproc
file offset length uri date hash ext copy ref_uri ref_date
warcrefs

Figure 2: WARCrefs deduplication process.

1. **Test the tools.** In deduplication, records are identified as duplicates and deleted, substituting in references to what is allegedly the original record. In the event that either the identification or the deduplication makes a bad move, this can result in data loss, where all copies of a resource are converted to **revisit** records, or where **revisit** records are set to point to something that is not a copy of the converted record. Moreover, as deduplication is data rewriting, data corruption is also a concern. If any of these errors occur, the damage will be irreversible once the rewriting is committed to both copies of the data. Therefore, extensive testing scenarios must be thought out and carried out before putting the tools to production use.

2. **Generate hash manifest.** The legacy web archive collection that was provided to the BA by the Internet Archive in San Francisco is in the old ARC File Format. Therefore, at this point, **warcsum** is to process ARC as well as WARC files.

3. **Convert ARC to WARC.** As the described deduplication process is designed to work with the new WARC File Format, conversion has to take place. The JWAT toolkit [12] will be used to convert one of the two copies of the data, leaving the other copy untouched as a fallback measure.

4. **Validate the conversion.** Generate a new hash manifest for the all-WARC copy of the data. Compare this manifest to the one generated pre-conversion. Investigate and act on discrepancies as necessary.

5. **Deduplicate.** Carry on with the deduplication process on the all-WARC copy of the data starting at stage 2 (**warccollres**). Be sure to use the post-conversion manifest, as record offsets and lengths in the WARC version of a file are different from those in the ARC version.

6. **Validate the deduplication.** Generate yet another hash manifest for the deduplicated data and compare to the post-conversion manifest to confirm non-duplicate records were not altered. Also, select a random sample of deduplicated records for testing through the access system. Investigate and act on issues as necessary.

7. **Commit to second copy.** If confident the deduplication process resulted in no damage to the data, commit the deduplicated set over to the second set that was kept unaltered throughout the process.

7. CONCLUSION

Deduplication is an effective technique for making smarter use of the storage infrastructure that supports a web archive, and also comes with a few desirable side effects, such as improving the quality of the collection. Proper identification of duplicates based on hash values and applying a second check to rule out collisions ensure the deduplication target is selected accurately. Best practices for standardizing how duplicates are represented in the WARC File Format have been drafted within the International Internet Preservation Consortium. The Bibliotheca Alexandrina is developing the tools needed to execute post-crawl deduplication of its web archive, and hopes to report on results and lessons learned from this petabyte-scale data rewriting job in a future venue. Other institutions involved in web archiving are welcome to put the tools to test on their own collections as well as contribute to improving the software.

8. ACKNOWLEDGMENTS

The Bibliotheca Alexandrina wishes to thank colleagues in the IIPC Harvesting Working Group, most notably, the National and University Library of Iceland, for work drafting the “Proposal for Standardizing the Recording of Arbitrary Duplicates in WARC Files.”

9. REFERENCES

[1] Btrfs Wiki. `https://btrfs.wiki.kernel.org/index.php/Main_Page`.

[2] Crypto++ 5.6.0 Benchmarks. `http://www.cryptopp.com/benchmarks.html`.

[3] Hash function. Wikipedia, the free encyclopedia. `http://en.wikipedia.org/wiki/Hash_function`.

[4] Heritrix Wiki. `https://webarchive.jira.com/wiki/display/Heritrix/Heritrix`.

[5] OpenSSL Wiki. `http://wiki.openssl.org/index.php/Main_Page`.

[6] OpenWayback. IIPC website. `http://netpreserve.org/openwayback`.

[7] Oracle Solaris ZFS Administration Guide. `http://docs.oracle.com/cd/E19253-01/819-5461/`.

[8] The BA web archive on GitHub. `https://github.com/arcalex`.

[9] Web Archiving. IIPC website. `http://netpreserve.org/web-archiving/overview`.

[10] Proposal for Standardizing the Recording of Arbitrary Duplicates in WARC Files. IIPC internal document, September 2013.

[11] M. Burner and B. Kahle. *ARC File Format Reference*, September 1996. `http://archive.org/web/researcher/ArcFileFormat.php`.

[12] N. Clarke. *Java Web Archive Toolkit (JWAT) Documentation*, October 2012. `https://sbforge.org/display/JWAT/Documentation`.

[13] J.-L. Gailly and M. Adler. *zlib 1.2.8 Manual*, April 2013. `http://www.zlib.net/manual.html`.

[14] J. A. Kunze, A. Arvidson, G. Mohr, and M. Stack. *The WARC File Format*, January 2006. `http://archive-access.cvs.sourceforge.net/viewvc/archive-access/archive-access/src/docs/warc/warc_file_format.html?revision=1.10`.

[15] M. Shibata. Web archives of devastated area sites and deduplication project. IIPC General Assembly presentation, 2011. `http://www.netpreserve.org/general-assembly/2011/Overview`.

[16] K. Sigurdhsson. Managing duplicates across sequential crawls. In *6th International Web Archiving Workshop (IWAW06), Alicante, Spain*, 2006.

[17] M. Stevens. Cryptanalysis of MD5 and SHA-1. Centrum Wiskunde en Informatica (CWI), Amsterdam, the Netherlands. `http://2012.sharcs.org/slides/stevens.pdf`.

Lessons Learned and Open Challenges Regarding Support for Data Management Plans and Research Data Management

Heike Görzig
University of Hagen,
Faculty for Mathematics and Computer Science
Universitätsstrasse 1
D-58097 Hagen, Germany
+49-2331-987-304
Heike.Goerzig@fernuni-hagen.de

Felix Engel
University of Hagen,
Faculty for Mathematics and Computer Science
Universitätsstrasse 1
D-58097 Hagen, Germany
+49-2331-987-304
Felix.Engel@fernuni-hagen.de

Holger Brocks
InConTec GmbH
Kirschenalle 7
D-96152 Burghasslach, Germany
+49-9552 931494
Holger.Brocks@incontec.de

Matthias L. Hemmje
University of Hagen,
Faculty for Mathematics and Computer Science
Chair of Multimedia and Internet Applications
Universitätsstrasse 1
D-58097 Hagen, Germany
+49-2331-987-304
Matthias.Hemmje@fernuni-hagen.de

ABSTRACT

This paper outlines an approach for developing tools and services that support automated generation, management, evolution and execution of data management plans (DMPs) by generating rules derived from the DMPs which can be applied to the data to be archived. The approach is based on existing models and tools that were developed in successive research projects SHAMAN, APARSEN, and SCIDIP-ES. The models include the Curation Lifecycle Model from the DCC, the OAIS Information Model and the Extended Information Model to support processes, domains, and organizations. An approach for deriving rules from policies is outlined to support using iRODS. OAIS and Context Information related to a data object is supported in a serialization using the OAI-ORE format.

General Terms

Frameworks for digital preservation

Keywords

SHAMAN, APARSEN, SCIDIP-ES, DMP, RDM, OAIS, OAI-ORE, data curation, automation, data management policies

1. Introduction and Motivation

In the Integrated Project SHAMAN that was funded by the European Commission in its Framework Program 7 (FP7) an *Archive-centric Information Lifecycle Model* (ACILM) had been introduced which conceptually supports pre-ingest and post-access activities by adding additional context information to Information Packages [1].

Building on this model and on related technical results of SHAMAN as well as on conceptual results from another FP7 project called APARSEN[2] a set of software tools had been developed in SCIDIP-ES[3]. The tools can assemble the required context and can package this context as provenance information together with the digital object itself as Information Packages, ready for submission, ingest, and archiving. Nevertheless, some remaining challenges regarding assembling provenance and authenticity information have been identified in one of the final reports of the project [4]. For example, higher usability of the preserved data can be ensured by establishing *Data Management Plans* (DMPs). These and related preservation policy processes ideally need to be defined at the beginning of a research project. In this way, preservation policies can be created much earlier than at production, assembly, and ingest time [4]. These preservation policies are then either created in isolation or in the context of an overall DMP. In many projects such DMPs are formally required which is, e.g., more and more the case in almost all public-funded research projects.

Funding agencies very often are requesting to make the research data generated in funded projects available for re-use in the future and therefore are demanding to elaborate DMPs already at proposal or at least at research-fund contracting time. To comply with this pre-requisite the DMPs have to include the archiving and preservation policy of the produced data of the project.

In order to maintain the archived data in an intelligible and interpretable way over a long period of time after the end of each such project, the generated data needs to be continuously enhanced with information about its production and usage context. The context to be preserved includes all known properties of the digital object and all operations carried out on it [5]. This includes the phases before ingesting the digital object to the archive and after accessing it. Within the preparation of a research project an initial production and use context can be foreseen and planned but during its execution the research process bears risks and uncertainties that can only be handled in a dynamic way when they appear. Therefore, on the one hand, DMPs describe the initial concepts in which the digital objects and their context need to be archived and preserved but on the other hand the DMP has to further evolve during the execution of the project. Therefore, also the initial production and use contexts and their related concepts have to evolve within the corresponding DMP.

Part of the production and use context is contained in the knowledge base of the designated community which can also change very quickly and unexpectedly [4]. Therefore, such context has to be identified, represented, added and maintained by three main actors in DMP context management. Typically, these actors are data producers, data managers, and data consumers. Adding DMP Context Information to data provenance information is usually a time-consuming, intellectual, i.e., human and manual process which is normally performed by the data managers. While working on this task, the data managers are also responsible for ensuring that the DMP's overall requirements are met. In large-scale projects and after the end of research projects the manual curation of this data might therefore become too costly or even impossible. Therefore, in order to achieve a more sustainable situation and working environment, the role of the data manager will have to be supported by appropriate tools and management processes. This means, that automating this work wherever possible has to become a goal of prime importance. To achieve this, automated curation and corresponding DMP support would have to incorporate all facets of context of the data object and respectively the evolvement of the context within activities of the data objects usage.

A DMP provides the concepts for archiving and preserving digital objects and also for preparing their potential re-use. It can be utilized to support automatic or at least semi-automatic contextualization. To support automation by means of applying Semantic Web technologies in this area of automation, any DMP needs to be supported by machine-readable semantic representations which are governed by an appropriate domain ontology. Within the context of our earlier work it has also been shown that preservation policy generation and DMP should be decoupled from the necessity to have knowledge of OAIS in order to support researchers in concentrating on the data in their field of expertise and scientific discipline. In this way, researchers should become free from the burden of having to know OAIS [4]. Therefore, a DMP can be seen as a dynamic document during a research projects life-time, it is evolving and needs to be adapted to changing needs.

In the following, we will first outline and analyze the requirements and challenges of the DMP domain in more detail in order to better explain the requirements and challenges of such an automated DMP support approach.

2. Overall Requirements and Challenges of Data Management Planning

As a basis for this identification of overall requirements and challenges, we will review the initial DMP of a very large research project that is funded by the European Commission: the so-called *Realizing an Applied Gaming Ecosystem* (RAGE) project that has just been kicked off and has made its DMP available to us for this initial analysis.

In *Research and Development* (R&D) projects like, e.g., RAGE, three roles or user stereotypes that are involved in *Research Data Management* (RDM) can be identified. These stereotypes span three dimensions that the DMP has to address. There is the *Formal Dimension* with project administration, the *Managerial Dimension* with project management and the *Operative Dimension* with project implementation and execution.

The Formal Dimension of DMP is spanned by the funding agencies' grant agreements (GA), corresponding laws and policies. The GAs usually provide the contractual framework for the DMP, specifying what the DMP has to accomplish and to comply with. Corresponding laws and regulations provide the legal, regulatory and policy-building framework. Alongside these contractual and legal specifications and requirements, corresponding DMPs have to be elaborated in compliance with all of them. In the case of our exemplar EU-funded project they have to follow the Horizon 2020 policies [6][7].

In the GAs, funding agencies mostly state that the DMP will, e.g., also have to comply with ethical guidelines, establish institutional and local procedures, specify the instruments for data collection, etc. The GA usually also refers to laws and regulations that will have to be fulfilled.

Project administrators in the back office usually study all these GA documents and corresponding requirements and challenges of the DMP specifications and have to extract a set of corresponding requirements and challenges and a corresponding representation schema of related constraints, targets, and activities which the project has to accomplish. For R&D project data access rights, duration of archiving, purpose of archival, sharing, and preservation policies according to the GA, policies and laws are formulated and specified. The Managerial Dimension uses the requirements and challenges schema to create the initial DMP.

To comply with the requirements and challenges created by the analysis of the formal DMP dimension, a RDM work plan is developed in the Managerial Dimension of the DMP. The RDM work plan describes the RDM scenario that has to be created 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. The implementation of the DMP is based on this RDM work plan. In the Managerial Dimension, quite often user stereotypes of a project coordinator, work package leader and task leader can be found.

The research project's R&D work plan is usually divided into work packages and is spread over various working groups. The work packages have organizational dependencies between each other; these can be dependencies on developed knowledge, results, deliverables, and experiences that will have to be shared between the working groups. These dependencies will be reflected in the work plan and will have to be defined in the DMP. Therefore, the creation of the DMP, e.g., needs to foresee communication and exchange strategies between the work package leaders. In analogy to the dependencies between work packages, there are also lower-level organizational dependencies on the level of tasks and activities within work packages. These tasks and activities will be carried out in working groups or other organizational entities within these working groups. In the R&D work plan the activities will have a time span assigned. In order to create the RDM work plan as part of the DMP, the project coordination has to work closely with the work package leaders, who are working together with the task leaders and so on. In each organizational layer of the R&D work plan activity that has to be performed, compliance with the GA and its corresponding DMP has to be achieved dynamically at the corresponding level of detail.

The creation of the DMP and its execution with the RDM work plan is a collaborative task. Between the work packages a consensus about dependencies, data management services and activities, needed sharing services and capacities will have to be achieved. Furthermore, the corresponding RDM will have to manage the *Intellectual Property Rights* (IPR) and corresponding access rights to project results and background and as well as data sharing policies in compliance with the constraints provided in the Formal Dimension. These IPR dependencies, access rights and sharing policies will have to be defined in the DMP and will have to be applied during RDM work plan execution when the data is finally generated, managed, archived and preserved.

The data are finally generated in the Operative Dimension, from this dimension the finest granularity of Context Information about the data to be generated will originate and will find its way into the RDM corresponding to the initial DMP.

The data producers who are, e.g., software developers and researchers in the project, form the Operative Dimension of DMP. Tasks and activities listed in the work plan are executed by them and thereby produce and use the data to be archived and preserved.

Staff working in this Operative Dimension contribute their specific input to the DMP and corresponding RDM activities. They will have the most concrete and operational information about the data to be produced and used and will be able to provide information about where the data is stored, data types, archive and file sizes, formats etc. Data generators will also be able to provide information about dependencies and relations between generated data. Information about relations between source code, binary code and application is also retrievable in this dimension.

In this dimension the produced and used data will have to be connected to its descriptive information in relationship to the specific knowledge of the R&D domain it has been produced for and used in. Therefore, produced and used data depends on the research domain, but also on other potentially related information already listed in the DMP. Deriving this knowledge from the input, e.g., researchers or developers have been providing in the planning phase will have to be added as descriptive information to the produced and used data.

After the digital object has been submitted, archived, and preserved, other users might later want to access and re-use the data and may add additional re-use Context Information to the digital object.

Information needed from the above described dimensions will have to be collected, managed, and finally packaged, ingested, archived, stored, and preserved. Therefore, respective tools will need to be developed. In order to introduce and analyze this overall set of problems in more detail, the related scientific challenges and technical requirements for these tools will now be described.

3. Scientific Challenges and Technical Requirements for DMP and RDM Support

The user interfaces needed for such support tools depend on the DMP dimension as well as on the user stereotypes, roles, and the type of activity users are performing.

The Functional Dimension has to create a validation schema against which the DMP can be validated. Administrators in the back office have to be enabled to formulate, e.g., IPR, access rights, storage requirements, archival, preservation, and sharing policies for data to be produced and used by the project. This schema is based on the GA as well as on corresponding laws, regulations, and policies.

Later, when the DMP is created in the Managerial Dimension, its validation has to be possible using the created RDM work plan as a schema and validation errors must be made visible. For creating the DMP in the Managerial Dimension, a first RDM work plan has to be developed. As in the R&D work plan sequences of activities and their dependencies are formulated in the DMP and its RDM work plan. Therefore, a set of interfaces is needed to support RDM activity creation and interlinking. The RDM work plan finally shall result in a valid, i.e., formally fully complying implementation of the DMP, resulting in depending on the project GA, different schemas for the RDM application and in different user interfaces supporting these processes.

In the Organizational Dimension different DMP and RDM user interfaces depending on the research domain will have to be created. The Organizational Dimension will also need access to the DMP and RDM interfaces where activities are created and edited. These DMP and RDM interfaces should allow the linkage of R&D domain data to R&D activities.

R&D data users and producers might have to add additional metadata to the digital objects. In addition, R&D data which has already been produced and used in another working group has to be accessible to potential "re-users".

As the creation of the RDM work plan that is complying with the DMP is a collaborative task, corresponding user interfaces for collaborative DMP and RDM activities are needed. R&D data producers have to coordinate with the R&D data users, when, who and what exactly has to be delivered. R&D as well as RDM tasks will have to be submitted to the data producers.

To describe the time schedule documented in the R&D, as well as in the RDM work plan, a sequential workflow needs to be modeled where the work packages are producing digital objects. A digital R&D data object is produced in a certain activity/task in a work package. This digital R&D data object might be needed as a resource in another activity/task. The digital R&D data object which will be a resource in an activity/task has to be produced in a preceding R&D or RDM activity/task, thereby creating dependencies between activities/tasks. As a consequence, the sequence of R&D and corresponding RDM activities/tasks, and as well as the dependencies between these activities/tasks will need to be expressed.

The activities/tasks in which the digital R&D data objects are created or used, will be performed by resources. These resources are part of an organizational structure. This organizational structure will be another part of the digital R&D data object's context information.

Finally, the activities/tasks in which the digital R&D data objects are produced as well as the digital R&D data objects themselves are specific to a certain R&D domain. In order to describe an activity/tasks and a digital R&D data object, R&D domain-specific vocabulary will be needed.

These different types of information will have to be combined in a way that the DMP and corresponding RDM can be adapted and maintained from this information. Furthermore, it needs to ensure that the digital R&D data object which was produced and used can be archived together with its production and usage context as provenance information. This has to be achieved in a sustainable way which allows automating future access and re-use activities.

4. Architecture, Data Modeling and System Distribution Challenges

Users, creating the DMP and the RDM work plan and producing and using R&D data are usually based in different locations within different organizations but they all need access to commonly produced, used, and shared R&D data. Part of the R&D data will be stable and will not change very much during the duration of the R&D project but especially in the planning phase of the R&D data production and usage collaborative work is needed and R&D as well as planning data will have to be interchanged very frequently. Depending on the user profile and roles, different DMP and RDM services and related data types and distribution models are needed.

There will also have to be different R&D, DMP, and RDM data types to be stored which are the digital R&D data objects and their R&D, DMP, and RDM context data. This data will have to be accessed by the DMP and RDM support tools. Some of the data will have to be stored in a central place but there are also others types of data that have to be submitted from a local system and later stored in the central system when they are ready to be uploaded.

The architecture of the system to support the creation and realization of DMPs and corresponding RDM work plans, needs to address the above mentioned challenges. For expressing the knowledge in DMPs and RDM work plans, an ontology and its vocabulary will have to be developed, as well as a schema that can support the creation of Information Packages based on this DMP ontology. As the development of a DMP and RDM involves actors of the three Dimensions, a structure for collaborative development and execution needs to be created, for example defining who can decide what in a DMP and how decisions are made.

Building on existing ontologies that represent activities in processes, domains and organizations, an ontology will have to be developed that combines these ontologies with the *Open Archival Information System* (OAIS, ISO 14721)[8] Information model for *Long Term Archival* (LTA) and *Digital Preservation* (DP).

On the basis of these DP models the respective user interfaces can be created. The system architecture will have to be created respecting the distributed and collaborative work, offering the mentioned features as a service. In terms of storage a local storage for active work and a centralized storage for archiving will have to be considered.

Policies described in the DMP will have to be formulated in a formal way to support the overall automating of the application of these policies within RDM activities.

5. Scientific and Technical State of the Art

Many funding agencies require the development of a DMP. The DMPs are very often part of the GA [9] [10]. The DMP aims to help organize the created data, by preparing storage so that created data can be submitted according to a planned procedure in order to find them when needed and can later be referenced. A DMP helps to maintain data integrity and avoid creating duplicates. DMPs also include archiving of information, which makes digital objects understandable and retrievable [9] [10].

There are different categorizations of the contents of DMPs. *Data Archiving and Network Services* (DANS)[7] identifies five [7]:

- Administration Information
- Data description
- Standards and metadata and everything else that is required to find and use the data
- Ethics and laws
- Storage and archiving

Information about time of collection and changes to the data also will have to be added. It might be necessary to justify the decision for a certain format, especially if it is a proprietary format, as, e.g., open access is in many funding agencies DMPs and corresponding policies required. It might also be expected to describe the relation and added value to existing data [9]. The sharing of the data might be restricted due to IPR, privacy concerns, or embargos. These restrictions will have to be outlined for the created data. For sharing and reuse of the data, information about which data will be shared with whom, who might be potential data users, it has to be stated when, how and where the data will be available and how the data will be licensed. Two aspects of data storage should be explained: short-term data storage, mostly locally, within the institution of the research project and long-term storage. For the later it needs to be explained, which data will be preserved, how the data will be preserved, including formats and technologies used. Budget and security issues might also be specified in the DMPs [11].

Many research institutions and funders are offering guidelines and templates for developing DMPs. More detailed help can be found in institutions that specialize in the development of DMPs. Some of these institutions do also offer some support tools for creating DMPs.

There are funding agencies that require periodical creation of DMPs, while others only request a DMP once [12]. Some funders ask for the DMP before the project starts, while others require the plan during project runtime. A DMP also includes information about how data will be managed and about policies to be applied. This will be discussed in the next sections.

The OAIS reference model is a widely accepted model for archiving digital objects. It consists of a functional model explaining needed functional entities to perform LTA and support DP. Furthermore, it provides an environment model describing involved actors which are data producers, consumers, and management, and it provides an information model for the structure of an Information Package that contains all data necessary to find, access, provide authenticity and the representation information to understand the archived data [8].

In OAIS, a digital object is interpreted using its representation information, by the so-called *Designated Community* (DC). The representation information itself is an information object and thus subject to representation information, the assignment of representation information is regressive until the assumed level of knowledge of the DC is reached. Over time the knowledge base of a DC can change, putting thereby the interpretability of a digital object at risk [8].

Parts of OAIS' functional model are the preservation planning functional entity and the access functional entity. The preservation planning functional entity supports recommendations and provides preservation plans to make sure that the information stored in the OAIS remains accessible and understandable over a long time to the DC [8]. The access functional entity provides services and functionalities that support users to discover, find and access digital objects.

Brocks et al. criticize the OAIS for leaving all responsibilities to what happens before digital objects enter an archive and after it leaves the archive to abstract stereotypes as producers and consumers. Important Context Information is not considered such as, for example, reviewing criteria in the process of scientific publishing [5].

The *Archive-centric Information Lifecycle Model* (ACILM) (Figure 1) developed in the project *Sustaining Heritage Access through Multivalent ArchiviNg* (SHAMAN) [1] can overcome this constraint and support the activities executed on a digital object during its life-span including the phases before and after archiving. The phases are creation, assembling, archiving, adoption, and reuse, where creation and assembling comprise the pre-ingest phase and adoption and re-use the post-access phase.

The creation phase involves a multitude of information describing, e.g., among other information the background of the data creation. In this phase so-called *Context Information* (CI) can be added to the digital object. The second phase when context is added to the digital object is the adoption phase, where the digital object can be re-contextualized; adding, for example, consumer information [5].

The creation of the digital object is based on the R&D work plan, the DMP and the RDM. In the assembly phase all information to meet the presumed needs of the designated community is assembled. In the archival phase policies concerning ingest, preservation and access are applied [1]. In the adoption phase the digital object received as an Information Package will be enhanced with process information as, e.g., representing examination, adaptation, and integration to enable understanding and re-use. The re-use of an object implies the dissemination and exploitation of an object and eventually transforms it or creates a new object. Adoption and re-use of a digital object can be subject to a research project's work plan and therefore underlay a set of research policies and rules. The OAIS information model has thus been extended.

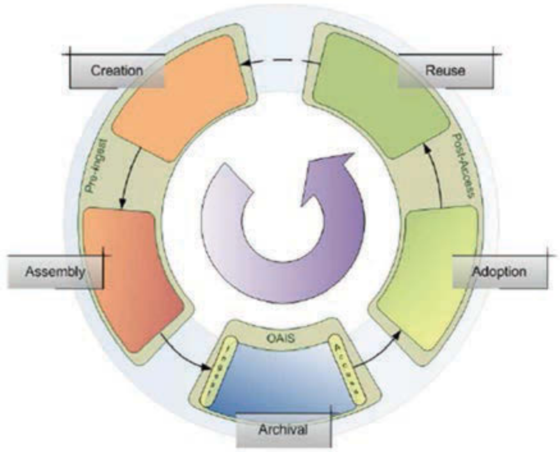


Figure 1 Information Life Cycle Phases[5]

The context of a digital object to be preserved over time comprises the representation of all known properties associated with it and of all operations that have been carried out on it. This implies the information needed to decode the data stream and to restore the original content, information about its creation environment, including the actors and resources involved, and information about the organizational and technical processes associated with the production, preservation, access and reuse of the digital object [5].

The context has been integrated into the OAIS Information Model without altering the concepts of its original information model [5].

The so-called *Extended Information Model* (EIM, see Figure 2) consists of the so-called *Context Information Package* (CIP) and the OAIS Information Package, sharing packaging information and package description. Additionally references exist to provenance, context and representation information.

Separating the context from the OAIS Information Package will allow for modeling the changes of concepts and terminology over time, characterizing production and (potential) reuse environments, and facilitates transfer to different communities by providing mappings of the underlying structured representations of concepts and relations [5].

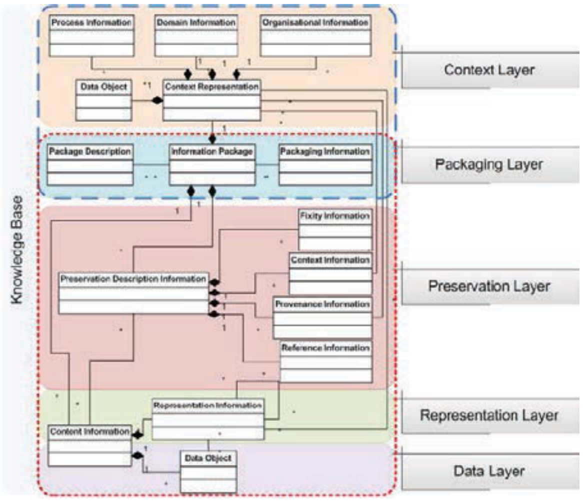


Figure 2 Extended Information Model [5]

The context representation consists of:

i) Process information
ii) Domain information
iii) Organization information

A context model has been created which can represent the information needed to describe the context of a digital object.

This model is based on the use of ontologies. The above introduced context dimensions i) can be represented by the use of domain ontologies, for ii) enterprise ontologies can be used and iii) can be described by process ontologies, where processes are divided in sequences of activities. The domain ontology defines the concepts and topics, but also their relations which are relevant for a particular application domain designated community. The enterprise ontology models the structural layout of organizational environments, such as affiliations, persons, or roles for describing a set of relevant concepts. The semantic classification of processes and activities as their building blocks requires their formal, hierarchical representation and description within an ontological structure [5]. Using the domain and the enterprise ontology rules can be specified as there are pre- and post-conditions, roles and interdependencies [13].

Brocks et.al explain the possibility of using OAI-ORE to develop ontologies that extend the OAIS information model in order to take into account Pre-Ingest and Post-Access processes much more than the OAIS suggests [5].

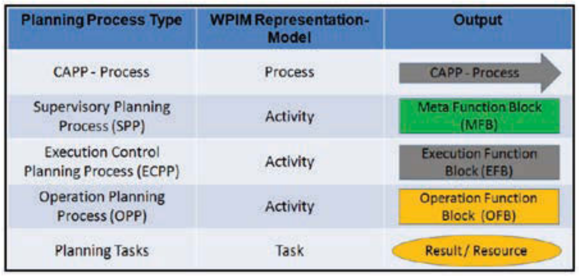
The *Extended Process Model* (EPM) integrates domain, enterprise and process ontologies into a conceptual unified process model [14]. This process model is meant to be applied in knowledge intensive processes with weakly structured activities, where the environment is dynamic and the process behavior and the entity concepts involved are unpredictable at design-time [14]. In this case traditional business process models with nearly static processes where the sequence of activities does not change frequently fail. The EPM is meant to enable flexible creation of processes, where a valid sequence of activities can be created by establishing rules for the activities by associating roles, objects, pre- and post-conditions and interdependencies [14]. The domain ontology comprises concept and topic information, the enterprise ontology can be used to describe roles and organizational structures and with the process ontology the dynamic aspects can be described [14].

To apply preservation management policies on digital objects, the policies will have to be described in a formal way. Therefore, the management policies will have to be refined in detailed policies which describe processes. For implementing these processes, procedures will have to be developed and described in workflows. These workflows can be formally represented in business process models/rules. For each refined policy each statement is described step by step by high-level rules in order to create a formalized description of the policy [1]. These high-level rules can later be transformed to operational rules, e.g., utilizing the *Integrated Rule-Oriented Data System* (iRODS) [15] for implementation. Using iRODS, small well-defined micro-services can be executed.

iRODS is open source distributed software to address key elements of data management. Rules derived from policies enable automation of data workflows, with a rule engine that permits any action to be initiated by any trigger on any server or client in the grid [15] and supports plug-ins for micro-services. iRODS micro-services can be executed based on these rules. The rules can e.g. initiate packaging operations using the Packaging-Service to create OAIS Information Packages for archiving or distributing access rights. iRODS can work in a distributed environment using a variety of storage locations and resource types. With an API it is possible to retrieve Data Objects from other storage applications [16].

The concept of *Knowledge-based and Process-oriented Innovation Management* (German: *Wissenbasiertes Prozess-orientiertes Innovationsmanagement, WPIM*) was developed to support capturing and usage of knowledge around innovation processes [17]. It assumes that innovation has both a knowledge and a process perspective which need to be used in combination. Activities of a process can be annotated with resources, such as experts and documents [17].

Gernhardt et al describe in [17] how WPIM and *Distributed Process Planning* (DPP) are used for supporting *Collaborative Production Process Planning* (CAPP) (Figure 3).



Planning Process Type	WPIM Representation-Model	Output
CAPP - Process	Process	CAPP - Process
Supervisory Planning Process (SPP)	Activity	Meta Function Block (MFB)
Execution Control Planning Process (ECPP)	Activity	Execution Function Block (EFB)
Operation Planning Process (OPP)	Activity	Operation Function Block (OFB)
Planning Tasks	Task	Result / Resource

Figure 3 CAPP Ontology based on WPIM Models and DPP Process Types and Resources/Results [17]

The overall CAPP-Process can be divided into three sub-processes (called activities) as there are the so-called *Supervisory Planning Process* (SPP), *the Execution Control Planning Process* (ECPP) and the *Operation Planning Process* (OPP) and finally the operational Planning Tasks as sub-processes of the three planning sub-processes. This means while the CAPP-Process is represented as one overall WPIM Process each CAPP sub-process is mapped to a WPIM activity and its operationalization is finally resulting in a set of tasks which implement the low-level Planning tasks within the three types of WPIM Activities corresponding to the planning dimensions. In the SPP of a CAPP *Meta Function Blocks (MFB)* are produced which represent generic information of process planning as there are e.g. machining technology and constraints [18]. The *Execution Function Blocks (EFB)* are created in the ECPP and can be seen as an instantiation of a series of MFBs; it includes scheduling information and monitoring events. In the OPP *Operation Function Blocks (OFB)* are produced. The EFBs get assigned to resources by means of the OCPP activity which outputs corresponding OFBs. In the OPP the OFBs are defined. These OFBs are directly linked to resources that execute these OFBs. To achieve a representation of this kind of sub-process structure on the basis of WPIM, the process planning levels ECPP and OCPP have to be represented as additional underlying WPIM activities of the same Master Process. Therefore the resulting outputs EFBs and OFBs of these processes have to be represented as planning results and therefore as knowledge resources that are handed over between these three planning activities [17].

6. Related Technical and Scientific Work

The *Open Archives Initiative Object Reuse and Exchange* (OAI-ORE) format described in [19] defines standards for the description and exchange of aggregations of web resources. A resource can be seen as a set or collection of other resources. This resource is called an aggregation. The resource map describes the relation the aggregation has to its aggregated resources. In other words an aggregation aggregates resources and is described by a resource map. The resource map must contain the aggregation it describes, enumerate the aggregated resources and may contain relationships between aggregated resources. In OAI-ORE RDF triples of subjects, predicates, and objects are used to formulate statements. For implementing OAI-ORE serializations with Java frameworks like, e.g., Apache Jena [20] and Protégé [21] have been created. In the SHAMAN project the OAI-ORE format was used first for defining an OAIS Information Package which has later been implemented in the SCIDIP-ES project.

The Packaging Service is using the OAI-ORE format for packaging. It has its origins also in the SHAMAN project and was implemented in the SCIDIP-ES project. It could be extended for serializing the above mentioned extended Information Packages. The Packaging Service is a web service which can receive requests for packaging OAIS Information Packages in zip archives containing a manifest file describing the Information Package. The manifest file can be serialized among others in OAI-ORE [3]. The Packaging Service can therefore support the archival phase of ACILM (see Figure 1).

A promising approach to support automation has been identified by means of the linkage of data objects to be preserved with their representation information using the so-called *Preservation Assistant* (PA)[4]. This approach will be used as a base for linking digital objects to their context. The PA originates from the same projects as the so-called *Packaging Service* (PS)[3]. It had been implemented to support data creators and managers to link data objects to archives with relevant information. A form is presented to the users, which they have to complete. On basis of this form the data to be archived will be automatically connected with the respective representation information [4]. The PA can therefore support the assembly phase of the ACILM (Figure 1).

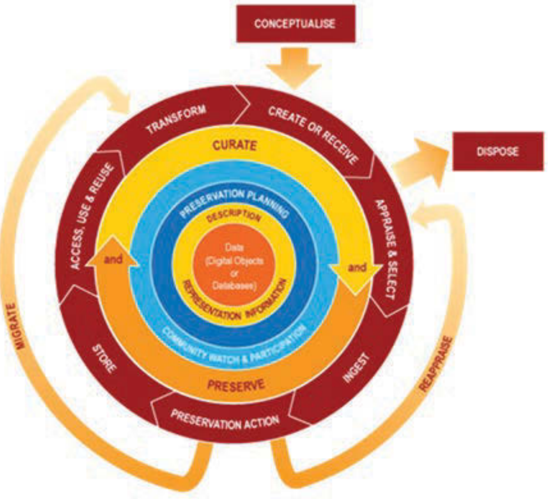


Figure 4 The DCC Curation Lifecycle Model [22]

In the *Digital Curation Centre* (DCC) the so-called *Curation Lifecycle Model* (CLM) was created, to provide a roadmap that ensures that all necessary steps in a curation lifecycle of RDM are covered [22]. As in the ACILM, the CLM of the DCC integrates activities before and after the preservation of the Data Object in its lifecycle model. While in the ACLIM the preservation of these activities is focused, in this lifecycle model the activities are specified for RDM.

Part of RDM is the curation of the created data. Digital curation involves maintaining, preserving, and adding value to digital research data throughout its lifecycle[23].

The RDM will have to interact during the different phases of a research project in the various steps of the lifecycle of a digital object. In the conceptualization phase, before the digital objects are created, capturing methods and storage will have to be planned. In this phase also requirements of the DMP will have to be incorporated, in order to comply with the funding agency's requirements. Assigning representation information, planning of preservation and curation will continue throughout the whole lifecycle of the digital object. Depending on the funder's requirements, DMPs will have to be created periodically throughout the lifecycle. The community will have to be watched and will have to participate, in order to *develop shared standards, tools and suitable software* [22]. So access, use and reuse of the digital object can be assured.

Two web-based approaches for establishing RDM the *Data Asset Framework* (DAF) and the *Collaborative Assessment of Research Data Infrastructures and Objectives* (CARDIO) have been developed by the UCC. The first is an interviewing tool covering main activities related to the curation lifecycle. The latter is a collaboration tool to find consensus by establishing RDM capabilities and finding gaps. The consensus is created by using ratings and comments [24]. Both tools are inspiring for creating user interfaces, but they themselves stay isolated in the RDM planning.

7. Modeling

In the same way as the CAPP-Process has been mapped to the WPIM-Process, the remainder of this paper will elaborate how WPIM and CAPP concepts can be applied in the next mapping step to the creation of DMPs. It should be noted that while CAPP was originally applied for planning processes in the manufacturing domain, it will now be applied to the RDM domain.

The three planning levels of the CAPP-Process comprise similar functionalities as needed by the three dimensions of DMP as described above. Therefore, to address and support the Formal Dimension of DMP, it would be necessary to execute a planning process like the SPP in CAPP where a first DMP on a meta-level is created. The activities of this process are, e.g., the formulation of requirements, constraints, organizational resources as well as target outputs. These meta-level DMP planning results are passed to the Managerial Dimension of DMP in a representation similar to a MFB. In the Managerial Dimension a planning process similar to the ECPP is needed in order to be able to express the DMP activities of this dimension including its inputs and outputs. This means that using the MFB input of the SPP the ECPP in the Managerial Dimension will define the DMP activities on the level of the RDM work plan. In this second level of the DMP planning process, which is now called ECPP, the first version of the DMP will be instantiated and responsible work package- and task-leaders need to create a corresponding RDM work plan. Activities of this process include the formulation of concrete entities, as there is the Process Information with its workflow and corresponding activities, tasks and dependencies, the Domain Information where the outcome (deliverable, knowledge, experience, result) of an activity in the Operative Dimension is described and the Organizational Information where the involved organizational unit and infrastructure is described. The output of the ECPP is representations similar to EFBs which will be handed over to the third level of DMP planning in analogy to the OPP which needs to be implemented in the Operative Dimension of DMP. This means that on this level the responsible actors have to concretely formulate the RDM operations that implement the RDM work plan. In other words, these types of activities need to be represented on the WPIM task level.

An EFB can either be directly assigned to resources in the Operative Dimension of an OPP activity which is producing the result of an OFB or they can be dynamically assigned at execution time on the level of the OPP. The results of an OFB are deliverables, knowledge, experiences and results, representing the OAIS Content Information. The OFB contains the most concrete and detailed information about the created results. The resources in the Operative Dimension are described in the Organizational Information.

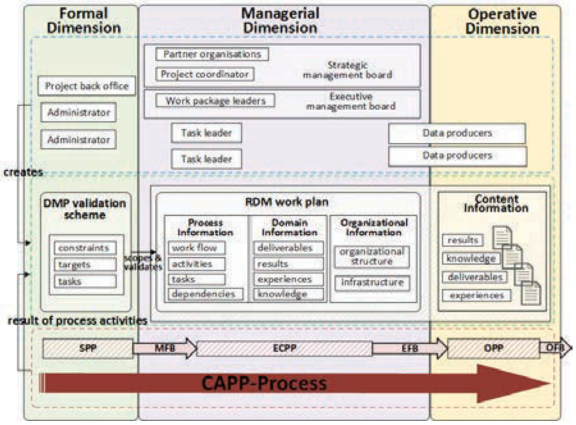


Figure 5 DMP Dimensions – CAPP-Process

Figure 5 displays a first draft for the design of the process models and information models in such a three-level DMP model that is inspired by WPIM and CAPP. The green ruled area represents the input for the OAIS Information Package with its extension to describe the context and provenance of a digital object in the Content Information, as described in ACILM and will be the information to be archived. The Process Information will be represented by WPIM-Processes which are structured in the CAPP-Process (Figure 5). The evolving concretized DMP will be extracted from the RDM work plan information.

For executing the DMP as mentioned above, iRODS rules could be used. They will have to be formulated by the Operative Dimension on the basis of an EFB/activity passed over by the Managerial Dimension. The formulated iRODS rules will have to be mapped against the policies formulated in the Formal Dimension. The policies are passed over to the Managerial Dimension in form of MFBs/activities. On the basis of these MFBs/activities, the Managerial Dimension has to be revised if the respective policy iRODS rules have already been defined. If this is the case the input data for executing the iRODS rules have to be selected and the EFB/activity can directly be passed to the iRODS rule engine for execution. Otherwise the rules will have to be mapped or formulated in the Operative Dimension. In this sense the iRODS rules can be seen as OAIS Content Information created by the Operative Dimension and will have to archived with its context.

8. Conclusions and Future Work

It has been outlined that for many of the remaining challenges starting points for research approaches do already exist. This includes modeling as well as technical challenges.

The CLM ACILM life-cycle models can guide activities and corresponding user-interfaces for creating OAIS-conformant information packages ready for ingest in an LTA.

The exiting tools and services for DMP creation and packaging are web-based to allow working in distributed environments. These web-based tools imply mostly the filling of forms that result from the DMPs in, e.g., a funding agencies' template. These tools mostly address the Managerial Dimension, needs for adding specific templates resulting from organizational backgrounds or research topic specific needs and thus affecting the Formal Dimension. The revised DMP planning tools normally do not allow for the delegation of work, as for example to the Operative Dimension for planning the concrete data to be created [25].

Modeling challenges for supporting OAIS can be approached using the EIM, which can be expressed using OAI-ORE and partly be serialized with the Packaging Service. Processes and organizations can be described and modeled using semantic models for enterprise resource planning and its application. Policies that give the context and explain the background of a digital objects creation, access and reuse can thus be formulated in a DMP in an ongoing research project.

In the modeling section an approach has been outlined using CAPP structure represented by WPIM to formulate DMPs respecting the three dimensions introduced at the beginning of this paper. As the analogies between CAPP and DMP have been shown, what remains is the formulation of an appropriate machine-readable representation of constraints as implied by laws, policies, regulations and contractual tasks. The function blocks of CAPP will have to be adapted to represent *Data Management Policy Rules* (*DMPR*) which will derive from the RDM activities represented by WPIM activities. Concrete instances of *Data Management Rules* (*DMRs*) could then be derived from the already rule-based DMPR representation in order to support an implementation using the *Integrated Rule-Oriented Data System (iRODS)* as an exemplar data management deployment infrastructure.

What remains is to formulate concrete representations of the DMPRs and DMRs.

Our future work can be divided into two subsets of R&D activities. The division into two subsets follows the suggestion in the lessons learnt from SCIDIP-ES[4] where the information modeling related to the direct users environment is separated from the OAIS Information Package creation. This means that users only have to deal with information of their research domain and does not need knowledge of the OAIS standard. The first subset consists of creating a concept of user interfaces that results in the creation a) of the DMP and b) formulating the rules that derive from the DMP. The second subset would use these rules for automating OAIS Information Package creation with Context Information by applying the formulated policies.

9. Acknowledgements and Disclaimer

This publication has been produced in the context of the RAGE project. The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644187. However, this paper reflects only the authors' views, and the European Commission is not responsible for any use that may be made of the information it contains.

10. References

[1] SHAMAN Consortium, 2011, Automation of Preservation Management Policies.

[2] APARSEN, APA | Keeping digital resources accessible, understandable and easy to find. [Online]. Available: http://www.alliancepermanentaccess.org/. [Accessed: 19-Apr-2015].

[3] SCIDIP-ES, 2013, D32 . 2 Generic Services / Toolkits and Robustness Research Report and Plan.

[4] SCIDIP-ES, 2011, D33 . 3 European ES LTDP infrastructure interoperability , architecture and governance model report.

[5] Brocks, H., Kranstedt, A., Jäschke, G., et al., 2010, Modeling context for digital preservation, *Stud. Comput. Intell.*, vol. 260, 197–226.

[6] European Commission, 2013, Guidelines on Data Management in Horizon 2020, no. December, 6.

[7] DANS, 2015, Datamanagementplan voor wetenschappelijk onderzoek, Den Haag.

[8] CCSDS, 2002, Reference Model for an Open Archival Information System (OAIS), *Forsp. Data Syst.*, no. January, 1–148.

[9] Jones, S., 2011, How to Develop a Data Management and Sharing Plan, *DCC How-to Guid.*, no. Dcc.

[10] MIT, Why manage & share your data? | Data management, *Online*. [Online]. Available: http://libraries.mit.edu/data-management/plan/why/. [Accessed: 08-Apr-2015].

[11] DATAVERSE, Data Management Plans — Dataverse.org. [Online]. Available: http://best-practices.dataverse.org/data-management/index.html. [Accessed: 10-Apr-2015].

[12] Doorn, P., 2014, Data Archiving and Networked Services World Wide Data Management : chaos of harmonie ? [Online]. Available: https://wiki.surfnet.nl/download/attachments/46794177/world wide data management ede - doorn - 10092014 .pdf?version=1&modificationDate=1410790253419&api=v2.

[13] Bayer, K., Kempf, S., Brocks, H., et al., 2006, A Multi-Agent Environment for the Flexible Enactment of Knowledge-Intensive Processes, *Cybern. Syst.*, vol. 37, 653–672.

[14] Brocks, H., Meyer, H., Kamps, T., et al., 2006, The Extended Process Model - Transforming Process Specifications into Ontological Representations, *Cybern. Syst.*, vol. 37, 1–6.

[15] iRODS, iRODS (integrated Rule-Oriented Data System). [Online]. Available: https://irods.org/. [Accessed: 18-Apr-2015].

[16] iRODS, 2014, iRODS Technical Overview. [Online]. Available: http://irods.org/wp-content/uploads/2012/04/iRODS-Overview-November-2014.pdf. [Accessed: 01-Jul-2015].

[17] Gernhardt, B., Miltner, F., Vogel, T., et al., 2015, Semantic Representation for Process-Oriented Knowledge Management based on Functionblock Domain Models Supporting Distributed and Collaborative Production Planning.

[18] Wang, L., Adamson, G., Holm, M., et al., Jul. 2012, A review of function blocks for process planning and control of manufacturing equipment, *J. Manuf. Syst.*, vol. 31, no. 3, 269–279.

[19] OAI-ORE, 2008, ORE User Guide - HTTP Implementation. [Online]. Available: http://www.openarchives.org/ore/1.0/http. [Accessed: 14-Apr-2015].

[20] Apache Jena, 2015, Apache Jena - Home. [Online]. Available: https://jena.apache.org/. [Accessed: 08-Jul-2015].

[21] Stanford University, 2015, protégé. [Online]. Available: http://protege.stanford.edu/. [Accessed: 08-Jul-2015].

[22] DCC, DCC Curation Lifecycle Model | Digital Curation Centre. [Online]. Available: http://www.dcc.ac.uk/resources/curation-lifecycle-model. [Accessed: 12-Apr-2015].

[23] DCC, What is digital curation? | Digital Curation Centre. [Online]. Available: http://www.dcc.ac.uk/digital-curation/what-digital-curation. [Accessed: 12-Apr-2015].

[24] Jones, S., Pryor, G., and Whyte, A., 2013, How to Develop Research Data Management Services - a guide for HEIs How to Develop Research Data Management, *Digit. Curation Cent.*, no. March, 1–22.

[25] Schoots, F., 2014, Datamanagementplannen Research Data Management Rapportage. [Online]. Available: https://wiki.surfnet.nl/download/attachments/42697239/Rapportage_DMP_Open.pdf?version=1&modificationDate=1397649373751&api=v2.

Developing a Framework for File Format Migrations

Joey Heinen
Harvard Library
90 Mt. Auburn St.
Cambridge, MA 02138
+1 (617) 373-3669
j.heinen@neu.edu

Andrea Goethals
Harvard Library
90 Mt. Auburn St.
Cambridge, MA 02138
+1 (617) 495-3724
andrea_goethals@harvard.edu

ABSTRACT

In this paper, we describe the development of a file format migrations framework at Harvard Library, using one migration case study, Kodak PhotoCD images, to demonstrate implementation of the framework.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice;

Keywords

Format Migrations; Migration Frameworks; Obsolete Formats

1. INTRODUCTION

As is well known to memory institutions, the act of preservation is never done, particularly once an object has been digitized. Digital material is just as susceptible to obsolescence as analog formats. There are a number of digital preservation strategies that can be employed in order to protect the usefulness of data, for example emulation, normalization and migration. Migration is chosen as a digital preservation strategy when the aim is to move content from its previously tenuous origins to a format with much greater promise in terms of support and usage [1]. Harvard Library uses migration as a primary preservation strategy because the Library's goal is to continue to provide networked access to digital collections on emerging platforms, without requiring researchers to physically come to the Library or to install special software.

Many institutions have demonstrated successful digital format migration projects (largely text-based) which focused on identification and preservation of significant properties within the format. However, few examples exist for how these projects can scale to a larger framework that can be continuously adapted for future format migrations, and for thousands or millions of files. At Harvard Library as a National Digital Stewardship Residency [2] project, one such possible framework was created. In order to test the viability of this generic framework, the project included the development of migration plans for three obsolete formats within Harvard Library's Digital Repository Service (DRS) [3] – Kodak PhotoCD, SMIL playlists, and RealAudio. While each format has its own challenges that will introduce deviations to a workflow, there are certain processes that will always be included in the migration workflow and plan. This paper does not diagram all aspects of the framework but outlines the main phases and components for creating a migration plan for a format. More information and documentation can be obtained by contacting Harvard Library directly.

Kodak PhotoCD is one of the first formats for digitized analog photographs and was used at Harvard largely for early photography and daguerreotypes collections. Real Audio and SMIL playlists were used for audio delivery at Harvard Library. These older formats are no longer deposited to the DRS but there is a great deal of content in the DRS in these legacy formats that needs to be migrated to modern formats. This project was made possible through the National Digital Stewardship Residency which allowed the resident (Joey Heinen) to develop this project over the course of nine months. Due to the time constraints of the project, the SMIL playlist project was only planned at a high level. The plan for the Kodak PhotoCD migration is complete though Harvard has yet to perform the migration.

2. CONTEXT FOR THE FRAMEWORK

The Harvard Formats Migration Framework is intended to inform the migration of obsolete formats regardless of the file format. While the specific format will necessitate variations to the overall framework, the framework will depict the general processes that must occur for each format in relatively the same sequence.

While the hope is that the framework can inform migration projects at large regardless of institutional context, there is obviously quite a bit that inevitably must be Harvard-specific – in particular, Harvard's organizational structure, policies, and digital preservation repository. Digital Preservation as a department resides within Preservation Services though also maintains strong ties with Library Technology Services (LTS), library-dedicated IT staff. Responsibilities for the Digital Repository Service (DRS) are shared between these departments with Preservation Services serving as the business owner and LTS as the technology owner. The DRS is both a preservation and an access repository. It provides Harvard affiliated owners with a set of professionally managed services to ensure the usability of securely stored digital objects over time.

There are a few DRS concepts that must be understood to understand the migration framework:

- A DRS **object** is a coherent set of content that is considered a single intellectual unit for purposes of description, use and/or management: for example a particular book, web harvest, serial or photograph.
- Each object conforms to a single **content model** which defines the object type (audio, still image, etc.). Content models define the supported file formats, object

structure, file and object relationships, roles and other key metadata.

- As defined by PREMIS "a **file** is a named and ordered sequence of bytes that is known by an operating system. A file can be zero or more bytes and has a file format, access permissions, and file system characterizations such as file size and last modification date." [4]

In a format migration plan, the files are the source for the migration and the plan will need to consider how to add migrated files to an existing object (which can contain different generations of the same files).

The associated content model of a format will also greatly affect the resulting migration plan. The content model affects which source file will be selected for the migration (the highest-quality version when possible) as well as how newly-migrated content is added to the repository, including how new relationships will be formed with the existing content. More complex formats may work interdependently with other files in order to produce the final results, such as the SMIL playlists which assist in delivering the RealAudio files. Understanding the content model, in particular the relationships that must be maintained or modified across migrations, is a crucial part of developing the migration plan.

2.1 Example – Kodak PhotoCD

Harvard Library preserves more than 7000 PhotoCD (PCD) files within the DRS and due to increasing difficulty in accessing (and thus preserving) these files over time, Digital Preservation Services decided to embark on a project to migrate these files to a modern target format. So that the Library would have a blueprint for conducting migrations for future obsolete formats, the PCD migration was used as a test case for developing the generic migration framework, noting the processes that must occur and generally in what order. While the overall generic framework had largely been designed as part of this project before testing it on PCD, it became an iterative process, updating the generic framework as experience was gained with this actual test case.

2.2 Related Projects

The project began with a literature review of migration projects. While an example could not be found of a format-agnostic migration framework that had been put to into production within an institution, many projects proved inspirational to the development of this framework, especially in the early stages. Workflow designs and models for building a migration plan from start to finish exist in the form of single projects/formats (National Library of New Zealand's WordStar to HTML4 [5], the Austrian National Library's TIFF to JP2 [6]) as well as larger institutional models for depicting roles and responsibilities (see Acknowledgements). Other projects demonstrated use of integrated tools to characterize/validate, convert, and QC migrated content (Austrian National Library). Others discussed use of registries and knowledge-bases to contain data on recommended tools and platforms for format migration (University of Illinois at Urbana-Champaign's Digital Preservation Interoperability Framework/Conversion Software Registry [7]) or to design and enforce holistic workflows and policies on migration (Technical University Vienna/AARIT's Plato [8], Norwegian University of Science and Technology's Multi-Criteria Decision Making model [9]). While these projects were not directly referenced in the design of the Harvard framework, the review helped to identify shared ideologies in what constitutes a successful migration and how to connect systems and technologies with theoretical processes.

3. THE FRAMEWORK

The specifics of this framework are much too large to describe here in detail, but the main components are stakeholder identification, migration workflow, and migration deliverables.

3.1 Stakeholder Identification

The identification of stakeholders first is deliberate – without clear roles and responsibilities, the migration project cannot start. Depending on the type of content, the particular departments and individuals may vary but the roles involved will stay somewhat consistent. The framework includes the following key stakeholder roles:

1. **Project Management** (those managing the overall migration project)
2. **Format Experts** (those who understand the format best)
3. **Content and Metadata Analysts** (those analyzing the content and metadata to be migrated and creating requirements documents and specifications)
4. **Plan Reviewers** (those reviewing plans and specifications)
5. **Systems and Technology Experts** (those helping to design the technical workflow and providing development and infrastructure support for the migration)

- **Content Owners** (curatorial stewards of the content to be migrated)

At Harvard, for format migrations, Digital Preservation Services plays the Project Management role, and serves as the primary Content and Metadata Analysts. The Format Experts vary, for the Kodak PhotoCD migration it is Imaging Services; for the SMIL playlists and RealAudio files it is Media Preservation Services. The Plan Reviewers include a variety of people across departments, and the Systems and Technology Experts role is played by Library Technology Services. The Content Owners vary depending on the content to be migrated, but will generally come from Harvard libraries, archives or museums.

3.2 Migration Workflow

The migration workflow can be broken down into five phases:

1. Plan for Test
2. Test
3. Refine Plan
4. Execute Plan
5. Verify Results and Wrap-Up Project

The workflow includes the creation of the migration plan as well as the actual migration. Each project phase can be further broken down into sub-phases and activities that may or may not produce deliverables.

- **Workflow Phases** are the five high-level parts of the migration workflow, each of which is further broken down into **Workflow Sub-phases** containing **Workflow Activities** (actions common to any migration)
- **Deliverables** include the migrated content itself, documentation or metadata generated along the way, diagrams, plans, or new revelations in digital preservation policies (e.g. storage and retention plans).

3.3 Migration Deliverables

The framework defines a set of deliverables for each phase, described here.

Phase 1: Plan for Test

- **Stakeholder Chart:** Identifies the departments and/or staff members who will fill roles during the migration project.
- **Format Specifications:** Where possible acquire authoritative descriptions of the relevant formats (formats to migrate but possibly also for the formats that will be migrated to)
- **Format Analysis Report:** Conclusions drawn from format technical specifications to determine significant properties, target formats, and possible conversion tools. Also include conclusions drawn from DRS metadata (or other relevant Harvard-specific sources).
- **Content Grouping Diagram:** Make-up of migration source files, their relationships to other files within an object, and the noteworthy technical attributes that will distinguish the ways that they are migrated (e.g. methodology, role, owner code, etc.). Includes useful SQL as an appendix
- **Target Formats/Conversion Tool Analysis:** Conclusions on target formats and conversion tools will be used in the test phase (and which ones will not), a scoring template which rates a tool/format's compliance with the defined significant properties of the format.
- **Migration Environment Specifications:** A list of necessary tools, plug-ins, and other software-based needs and the necessary OS/platforms/processors needed to support the software. Consider short-term storage capacity needs if necessary.

Phase 2: Test

- **Testing Conclusions Report**: Findings of the tests, the testing parameters, metrics for determining acceptability of the conversion, analysis of embedded metadata, and decisions on the best courses of action for the migration.

Phase 3: Refine Plan

- **Migration Pathway Diagram:** How migration will be performed based on content sub-groups, how migrated files will be created relative to conversion tools and custom settings, target formats and how these will be deposited and related to existing files in the DRS.
- **Migration Workflow Diagram:** Workflow processes mapped against RACI model (roles broken down into Responsible, Accountable, Consulted, and Informed) for stakeholder involvement. The workflow is broken into the 5 migration phases. Within each RACI grouping, define the plan components (see Format Migration Framework section). Uses shapes to correspond with the action (Process, Sub-process, Consensus/Decision, Changes to Content, Conditional Factors).
- **Migration Plan:** This is a comprehensive summary of all conclusions drawn from analysis and testing. Emphasis will be placed on necessary tools and systems for grouping, converting, and ingesting files based on content groupings.
- **Metadata Mock-up:** A wishlist for augmentation metadata to include information about the migration (processes, tools, etc.), generally for recording migration-specific PREMIS events.
- **Batch Ingest Mock-up:** A step-by-step process of how batches will be created based on migrated content grouped along with existing files within an object.

Phase 4: Execute Plan

- **Migration Checklist:** Record of the migration process, including key staff involved and tools used

Phase 5: Verify Results and Wrap-Up Project

- **QC Report:** Record of the verification of the converted files (passes based on decided metrics through QC tools if available).
- **Migration Conclusions:** Summarize lessons learned noting any anomalies or adjustments made along the way that might help to inform modification of framework or plan documentation.

3.4 Migration Workflow Example: Kodak PhotoCD Images

In this section, each phase and sub-phase of the generic framework is briefly described and then followed by an illustration from developing the PCD migration plan.

3.4.1 Phase 1: Plan for Test

3.4.1.1 Sub-Phase 1: Project Start-Up

Project start-up involves identifying key stakeholder roles and responsibilities, and setting the stage for the analysis and planning which must imminently take place. Acquiring technical specifications and format reports, exploring the provenance of the format within the institution/collections, and securing a technical environment for performing the basic analysis are all essential first steps. Additionally, it is essential to identify parallel library projects that will affect the migration from the outset so that they are embedded within the plan.

For the PCD plan, staff in Imaging Services served as Format Experts and participated with others as Plan Reviewers. They helped to analyze the significant properties of the format and to design the testing environment. Library Technology Services would be responsible for Systems and Technology Experts. Digital Preservation would be responsible for Project Management and Content and Metadata Analysts.

A project to migrate all metadata from an older version of the DRS to a newer version was running concurrently to the development of the PCD plan. The metadata migration project made changes to the content model associated with PCD images (most importantly in how file-to-file relationships are described). This metadata migration project was considered at many steps of the PCD migration plan.

3.4.1.2 Sub-Phase 2: Analysis

The first steps of the analysis sub-phase are to research the format specifications to identify the important technical characteristics of the format and to analyze the DRS metadata to break content down into groups relevant to the migration. This analysis should result in an early understanding of what the requirements might be for target formats and tools (for conversion, metadata extraction, and so on). Naturally, the technical characteristics and ways that content can be broken down will vary considerably based on the format, but this generic component will always be a necessary precursor to developing a format migration plan.

Kodak PhotoCD is a proprietary format that was popular in the late 1990s as a means of creating digital surrogates of analog photographs and slides. While it was at first adopted as an archival format it was eventually noted that its use of proprietary rendering software and applications as well as its unique forms of compression and color encoding were contributing factors in the format's eventual obsolescence (broadly discussed as early as 2005) [10].

Based on analysis of various technical specification documents, web forums, and white papers, the following significant properties were defined for the PCD format:

- PCD used PhotoYCC, a unique color space for segmenting luminance information from two chrominance channels (third channel interpolated) and to encode color information that goes beyond that which is conventionally contained within 255 decimal values. The color space is device independent meaning that it is designed to be rendered on any number of output devices (both analog and digital) [11]. Few file formats can support and render this color space.
- The PCD format supports a number of Scene Balance Algorithms (SBAs) (used for automatic lightness and color-balance adjustment) that can be applied at the time of scanning. This means that for different photo stocks and materials, different algorithms could be used to encode the scanning data to account for nuances in light and color [12]. SBAs need to be understood and accounted for by a conversion tool so that color (chrominance) and light (luminance) are presented as accurately as possible.
- In addition to technical metadata, provenance metadata from the digitization and from the disk encoding process history can be found embedded within the file.
- Image Pac compression, used by the PCD format, is a very efficient form of mathematically lossless encoding which may not be compatible with reporting or conversion tools. For example, ImageMagick [13] reads the Harvard PCD files as 768 x 512 (the Base image rather than Base16) and reports the compression scheme as "undefined." It will be important that the migration tools know how to unpack the images and read them at their fully uncompressed resolution (2048 x 3072) [14].

The DRS metadata is stored in an Oracle database. For DRS analysis, the database needed to be queried using SQL in order to explore the metadata looking for key technical and historical differences among the content as well as the relationships between content in this format and other formats. The results of these queries and analysis of the data is expressed pictorially in the **Content Grouping Diagram.** The most useful metadata for classifying the PCD files into groups was found in the methodology field, which is where free-text narratives described the digitization process for the file. This metadata was used to group the files into three essential groups based on their collections – The Harvard Daguerreotypes, the Horblit Collection, and the Richard H. Ree Collections. The first two collections feature early photography holdings (mostly daguerreotypes) that were some of the first photo digitization projects at Harvard in the late 1990s. They both employed the Kodak PhotoCD scanning process though used a unique Scene Balance Algorithm to account for different photo stocks that were used to initially photograph the images objects. The process used to create the Ree Collection is a little less clear cut, especially given this line from the methodology statement associated with this content:

"Ree's PhotoCD format images were processed using Adobe Photoshop 6.xx and 7.xx. The PhotoCD files were imported into Photoshop as 16 bit RGB TIFF files using the built-in import module with the "universal E-6" film term. Each image was individually processed to compensate for any obvious color casts and to achieve, to the extent possible, natural tone and color."

It is not noted how the images were digitized, simply that they were imported as digital. It also seems that images were individually corrected at the discretion of the Imaging Technician such that a monolithic film term setting wouldn't help to account for any of the original color or light settings (even if "universal E-6" was used to import the images into PhotoShop). Unfortunately no other provenance documentation exists from the original digitization or deposit of the digital images so the best that can be done is to analyze additional metadata within the DRS (and also to keep a sharper eye on this collection during conversion testing).

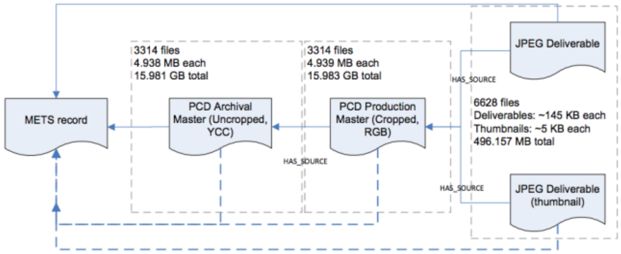


Figure 1: Content Grouping Diagram for a Still Image object from the Horblit Collection. This particular grouping shows PCD as both an Archival Master (Uncropped) and Production Master (Cropped) which will both be used as migration sources.

Additional DRS metadata was helpful for designing the **Migration Pathway Diagram** for each set of files that could be migrated as a group. "Roles" metadata defines the file's placement within the production workflow, namely for the Still Image content model if the file is an Archival Master, Production Master, or Deliverable. This was useful for determining which file to use as the source for the migration. For the Horblit Collection, PCD was used for both Archival and Production Masters with JPEGs as deliverables. The Archival Masters were fully uncropped including color bars for calibrating the scanning equipment to the imaging environment. The Production Masters were cropped and used to generate deliverable JPEGs for the web. It was decided that new Archival and Production Master images would need to be generated during migration. For the Harvard Daguerreotypes and Richard Ree collections, a PCD Archival Master (cropped) was used to generate a TIFF Production Master, which was used as the source for generating JPEG deliverables. It was decided that the TIFFs would be removed since they were generated using inferior PCD conversion software that did not account for SBA settings. For this case the PCD Archival Masters would be used to generate both a JP2 Archival and Production Master.

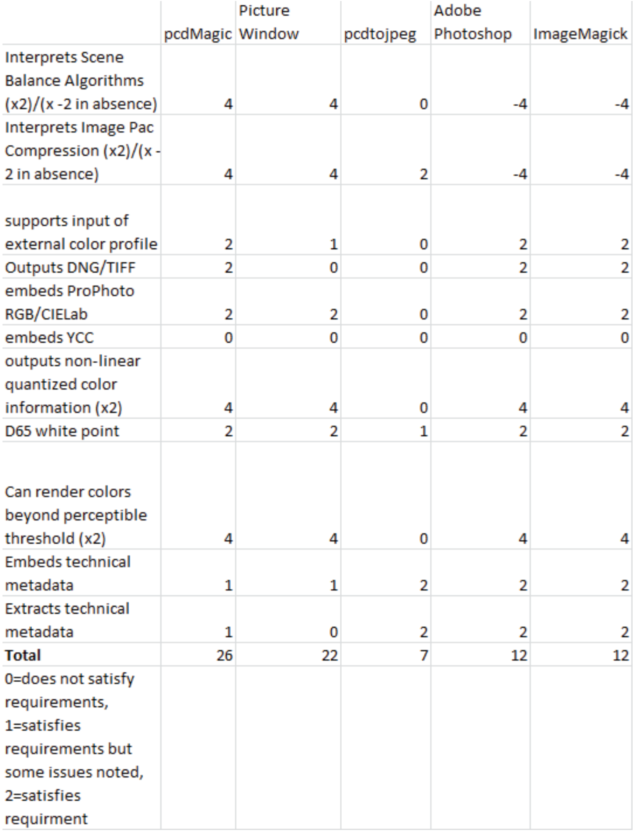
Other metadata was also useful in building the overall framework and migration plan, but in unexpected ways. The analysis uncovered errors in manually-submitted metadata for some files, specifically for metadata about Color Space, Compression, Dimensions, Scanning Systems, Vendor/Producers, and Roles. For example, the Production Masters in the Horblit Collection were all listed incorrectly as RGB Lossless images instead of YCC Image Pac. This would need to be corrected before creating the final batches for DRS ingest. Additionally, all the images from the Harvard Daguerreotypes that should have been listed as Archival Masters were marked as Production Masters, metadata that would also need to be corrected before DRS ingest (or even before execution of the migration in the event that scripts are used to pull PCD images from the DRS based on their “Role”).

3.4.1.3 Sub-Phase 3: Confirming Migration Criteria

This sub-phase includes working closely with the Format Experts to confirm the analysis results - that the format’s significant properties have been defined, the results of the metadata analysis look correct, and that there is clear criteria for choosing among conversion tools.

In the PCD case, the ideal target formats which emerged out of the analysis phase were confirmed by Imaging Services staff. They confirmed that the YCC color space is not supported by many image formats but can be mapped to ProPhoto RGB with minimal-to-no loss in information [15], CIELab also demonstrating good results [16]. Of course, all of this would be inconsequential if there were no available tools for performing these conversions, leading into an analysis of the available tools.

A scoring system as shown in Figure 2 was used to compare conversion tools as well as for choosing possible target formats (not pictured), resulting in the **Target Formats/Conversion Tools Analysis** report. A score was applied to each of the criteria (based on the defined significant properties of the format) which for some criteria involved a weighted score. In some instances an especially important criteria could incur a negative fee if the tool did not support this feature, meaning that the tool in general was not sufficient for use in the migration. In scoring tools based on their ability to meet the needs of the format migration and adding up the scores to generate a final value, it was much clearer to see which tools would generate a more desirable outcome, and especially which tools were unacceptable and would not require inclusion in the actual migration testing.



	pcdMagic	Picture Window	pcdtojpeg	Adobe Photoshop	ImageMagick
Interprets Scene Balance Algorithms (x2)/(x -2 in absence)	4	4	0	-4	-4
Interprets Image Pac Compression (x2)/(x -2 in absence)	4	4	2	-4	-4
supports input of external color profile	2	1	0	2	2
Outputs DNG/TIFF	2	0	0	2	2
embeds ProPhoto RGB/CIELab	2	2	0	2	2
embeds YCC	0	0	0	0	0
outputs non-linear quantized color information (x2)	4	4	0	4	4
D65 white point	2	2	1	2	2
Can render colors beyond perceptible threshold (x2)	4	4	0	4	4
Embeds technical metadata	1	1	2	2	2
Extracts technical metadata	1	0	2	2	2
Total	26	22	7	12	12
0=does not satisfy requirements, 1=satisfies requirements but some issues noted, 2=satisfies requirment					

Figure 2: Scoring acceptability of conversion tools

Based on this analysis Digital Preservation and Imaging Services decided that pcdMagic was the best conversion tool based on its ability to meet all of the essential criteria including its ability to interpret YCC, SBAs, and Image Pac compression. It could output both TIFF and DNG with a ProPhoto RGB color space. Additionally, it can accept external color profiles for more precise rendering of the color and light information (as an alternative to SBAs). Fortunately, Imaging Services owned color profiles that were specific to film terms used in some of the original PCD scanning software, something that would prove to be a boon to a successful migration plan.

3.4.1.4 Sub-Phase 4: Metadata Analysis

This sub-phase is an exploration of the tools that can best best characterize the format and/or provide process history information about the conversion process.

Though pcdtojpeg was found to not be an ideal tool for converting the format, it was the best at outputting provenance metadata about the file (scanning information, SBA settings, etc.). ImageMagick, another rejected conversion tool, was also a good metadata extraction tool because it was able to extract Exif metadata and technical metadata about the RGB channels. Exiftool was also used for metadata analysis, particularly for analyzing the images post-conversion. In the Exiftool output, the DNG files produced from the PCD files would present a color space of “pcdMagic DNG Profile” under “Profile Name” whereas the TIFF files would present a color space of “ProPhoto RGB” under “Profile Description.” This led to a decision to choose TIFF as an intermediate output during the conversion because the color space is more standardized and a better choice for preservation.

3.4.1.5 Sub-Phase 5: Moving Into Test Phase

At this point in the workflow the tools and target formats for the conversion have been decided; this sub-phase includes additional testing to determine some of the conversion details including how the tools would be run and any tool parameters.

For the PCD plan, this mostly came down to the Scene Balance Algorithms and how to most accurately depict and capture color information from the image. The environment for perfoming the migration was determined, which in this case was a Mac OS X environment (the most recent release of pcdMagic works with the OS X environment and had no additional dependencies besides an optional addition of color profiles in the ColorSync folder). pcdMagic is available for both Mac and Windows platforms however the Mac version is the only version that allows for external color profiles. For the testing phase a test laptop was used knowing that it would be possible to transfer the license for the tool to a production workstation when ready to move to Phase 4: Execute Plan.

3.4.2 Phase 2: Test

3.4.2.1 Sub-Phase 1: Create Sample Conversions

In this sub-phase a representative subset of the content is converted in preparation for analyzing the results together with the Format Experts.

For each PCD content grouping (determined by the methodology/collection with which the image is associated), 6-8 images were selected for testing. All five of the Kodak color profiles provided by Imaging Services (Color Negative, 4050 E-6, 4050 K-14, Universal E-6, Universal K-15) and a sampling of the general settings provided within the tool were tested (largely for comparison to demonstrate inadequacy of the pre-existing settings). The images were output in both TIFF and DNG format using various settings in order to determine the more ideal target format.

3.4.2.2 Sub-Phase 2: Assessment of Sample Conversions

In this sub-phase a combination of manual and automated tasks are performed with input from the Format Experts to make final decisions about how the migration will be performed and to verify that the conversion will be acceptable.

The PCD test conversions were viewed within PhotoShop. For additional comparison, multiple film terms were selected for each image in both TIFF and DNG formats. After refreshing the images, RGB histograms were consulted to make sure that no clipping of information had occurred and to see which images produced the widest gamut with the most evenly distributed waveform throughout. In some instances Imaging Services staff would make final judgments based on visual appearance, determining which images presented the best real-world results (not overcompensating in any of the RGB channels). The key decisions as documented in the **Testing Conclusions Report** were:

1. For the Horblit Collection, the Kodak Color Negative film term produced the best results. This was commensurate with the methodology statement.
2. For the Harvard Daguerreotypes, the 4050 E-6 film term produced the best results. This was commensurate with the methodology statement.
3. For the Richard H. Ree Collection the 4050 E-6 film term generally produced the best results (though in some cases was not as definitive). This is not commensurate with the methodology statements which said that the Universal E-6 film term was used though this does not appear to be the case. It will be necessary to decide if this group will require a more detailed conversion process where all 177 images are converted with their own unique settings.

During this process no discernible differences were seen between TIFF and DNG outputs (confirmed by subtracting pixel information from images and also comparing histogram readings) and that cropped and uncropped versions of the same image produced virtually identical color mappings (with the exception of borders and presence of color bars). However, as noted earlier, the color space associated with DNG was less preferable to the ProPhoto RGB found in the TIFF output.

As an extra step of quality control, characterization tools were used to ensure that embedded metadata was not lost (largely provenance).

3.4.3 Phase 3: Refine Plan

3.4.3.1 Sub-Phase 1: Analysis of Systems in Place

In this sub-phase the migration plan, which up until this time has been largely theoretical, is integrated with the Harvard Library infrastructure. Decisions need to be made about how the DRS files relate to the files that will be produced in the migration, and how the files produced in the migration will be integrated into existing DRS objects, and which files will be retained.

In order to gain insight and approval from all relevant stakeholders, the migration process is expressed pictorially in the **Migration Pathway Diagram (See Figure 4 for a PCD example).** The overall process employed for the entire PCD format migration (including initial planning and analysis phases) is expressed in the **Migration Workflow Diagrams** along with stakeholder roles and responsibilities. A narrative version is outlined in the **Migration Plan** document.

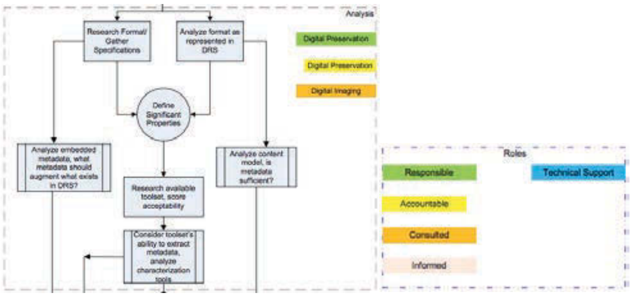

Analysis
Research Format/ Gather Specifications
Analyze format as represented in DRS
Digital Preservation
Digital Preservation
Digital Imaging
Define Significant Properties
Analyze embedded metadata, what metadata should augment what exists in DRS?
Analyze content model, is metadata sufficient?
Research available toolset, score acceptability
Consider toolset's ability to extract metadata, analyze characterization tools
Roles
Responsible
Technical Support
Accountable
Consulted
Informed

Figure 2: Excerpt of Migration Workflow Diagram with example of RACI color-coding for Stakeholder involvement

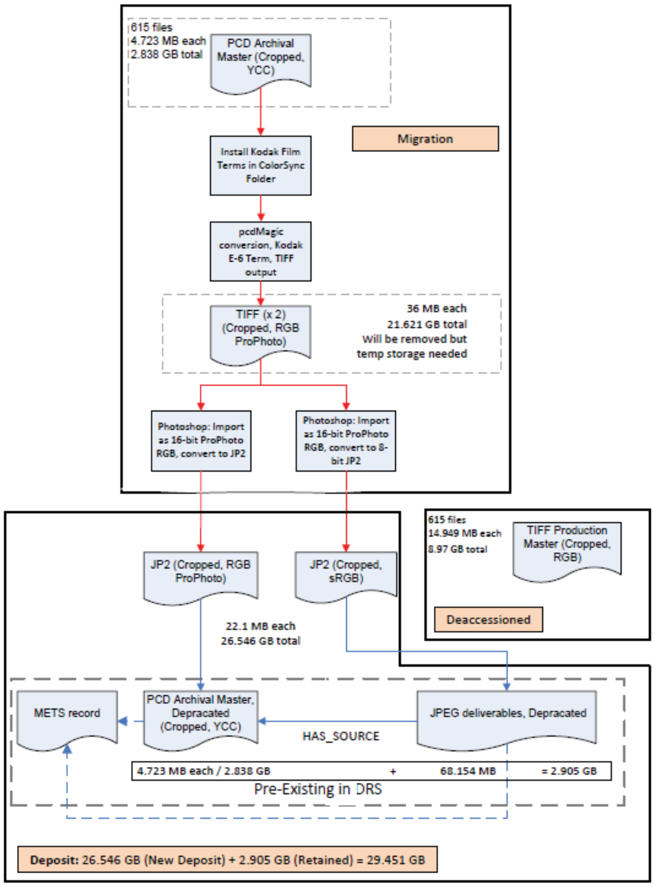

615 files
4.723 MB each
2.838 GB total
PCD Archival Master (Cropped, YCC)
Migration
Install Kodak Film Terms in ColorSync Folder
pcdMagic conversion, Kodak E-6 Term, TIFF output
TIFF (x 2) (Cropped, RGB ProPhoto)
36 MB each
21.621 GB total
Will be removed but temp storage needed
Photoshop: Import as 16-bit ProPhoto RGB, convert to JP2
Photoshop: Import as 16-bit ProPhoto RGB, convert to 8-bit JP2
615 files
14.949 MB each
8.97 GB total
TIFF Production Master (Cropped, RGB)
Deaccessioned
JP2 (Cropped, RGB ProPhoto)
JP2 (Cropped, sRGB)
22.1 MB each
26.546 GB total
METS record
PCD Archival Master, Depracated (Cropped, YCC)
JPEG deliverables, Depracated
HAS_SOURCE
4.723 MB each / 2.838 GB
+
68.154 MB
= 2.905 GB
Pre-Existing in DRS
Deposit: 26.546 GB (New Deposit) + 2.905 GB (Retained) = 29.451 GB

Figure 4: Excerpt of Migration Pathway Diagram for the Harvard Daguerreotype/Richard Ree Collection. The top box shows the migration process that will start with a PCD Archival Master, produce an intermediary TIFF image, and then will be converted to two JP2 images. The bottom box shows that the two JP2 images will be deposited to the DRS and related ot existing DRS images. Migration processes are shown as red lines; documented DRS metadata relationships are shown as blue lines.

A most essential final step in this phase is to finalize the definitions of the migration environment – the computational systems, storage processes (both temporary and permanent), key hand-offs of content throughout the workflow and tool set-up and use requirements. This list will help to expose any additional development that may need to take place on the existing technology. These needs should be considered in the overall Migration Plan and Workflow Diagram though are detailed specifically in **Batch Ingest** and **Metadata Mock-Ups**.

For the PCD plan, it was decided that in all cases two JPEG2000 JP2s would be created and would replace previous master and derivative files, one with the RGB ProPhoto color space to serve as an Archival Master (PCD as back-up), and one with the sRGB color space to serve the purposes of both Production Master and Deliverable. It is worth noting that while a TIFF is generated from the pcdMagic tool that the JP2 is the ultimate target format that will be saved from the migration. In the case of the Harvard Daguerreotypes and Richard Ree projects, the cropped PCD Archival Master would serve as the source for both the ProPhoto RGB JP2 and the sRGB JP2. For the Horblit Collection the uncropped Archival Master would serve as the source for the ProPhoto RGB JP2 and the cropped Production Master would serve as the source for the sRGB JP2.

Upon ingest of the JP2 files, new relationships will need to be added to link the JP2 images to the source files that they are replacing. The TIFF Production Masters from the Harvard Daguerreotypes and Richard Ree Collections may not be retained since as described earlier they were generated using inferior software that did not account for the SBA settings. Though new JP2 deliverables are being created as part of the migration, the older JPEG deliverables need to be retained since they have persistent names (URNs) published in catalogs and web pages. The original PCD images will be kept in the unlikely event that a future migration effort is performed (with newer, better tools on the market, which is also highly unlikely). The PCD images are relatively small so they do not affect storage capacity too greatly.

3.4.4 Phase 4: Execute Plan

3.4.4.1 Sub-Phase 1: Schedule Migration

In this sub-phase the migration project is scheduled and staff resources for the migration execution are assigned.

At the time of this writing the PCD migration has not been scheduled yet. This project was being done as an NDSR residency project, and the residency term ended after nine months, putting the project on hold. The project remains a high priority but will have to wait until there are staff resources within Digital Preservation Services that can continue this work as this department is taking the lead on the project.

3.4.4.2 Sub-Phase 2: Custom Development

Especially for the first migrations within an organization, they will likely require custom development by the Systems and Technology Experts. In some cases new scripts will need to be created to create a migration pipeline in which conversion tools can be inserted and removed as needed, in other cases existing tools will need to be modified.

In the PCD case Library Technology Services will need to modify its DRS ingest tools to be able to add the files created through the migration to the existing DRS objects. The existing DRS deposit tools can only add new objects to the DRS, not modify existing objects. This is indeed an issue since some products of the migration will replace content previously contained within the image object (e.g. TIFF intermediate files that were created using inferior conversion tools/processes which will be replaced by new JP2 files). In addition they will create a script so that pcdMagic can be called programmatically.

3.4.4.3 Sub-Phase 3: Conduct Migration

This is the sub-phase where the actual migration is conducted. It concludes exporting the content that will be used as the source of the migration to a temporary storage area, conducting the migration according to a **Migration Checklist**, and depositing the content to the DRS.

In the case of the PCD migration, the source PCD images and their associated METS metadata files will be exported by Library Technology Services to a directory structure specified by the Analyst. The values of specific metadata elements (methodology, role, and relationships) will be used by the migration tools to know which parameters to use and which files to create.

3.4.5 Phase 5: Verify Results and Wrap-Up Project

After the migration and ingest to the DRS there will be need for the final checks, documentation and clean-up. The metadata and reports that are generated throughout the workflow should be re-checked to confirm the success of key processes, that the migration was complete and that the metadata and content results are as expected and documented in the **QA Report**.

This is also the sub-phase where the de-accessioning plan developed earlier in the workflow is revisited to see if additional steps need to be taken, for example if files should be deleted or simply made inactive. This is also the appropriate stage for reviewing all the documentation produced throughout the migration. Ensuring that each document accurately reflects the final process is very important as these will likely be referenced for future migration projects as well as serving as authoritative provenance documents for demonstrating the chain of custody of the content. At this point it should be decided if any of these documents merit inclusion in the repository along with the files. The framework ends with writing any lessons learned in a **Migration Conclusions** document to inform future migrations.

4. CONCLUSIONS

While the Kodak PhotoCD and RealAudio/SMIL Playlists migration plans are still underway, simultaneous development of the plan for each format and the generic migration framework has helped to conceptualize the process for each format, identify aspects common across the format plans, and provide more certainty that a generic framework is possible. While the framework is very specific to the processes and procedures at Harvard Library, it is hoped that the framework will be helpful to other institutions as they approach migrations as a preservation action for their digital collections at scale.

5. ACKNOWLEDGMENTS

We would like to thank David Pearson of the National Library of Australia and Peter McKinney of the National Library of New Zealand, both of whom generously shared drafts and finalized versions of documentation and workflows from their format migration projects. We would also like to thank the Institute of Museum and Library Services for making this project possible through the funding of the NDSR Boston residency project. After finishing his residency at Harvard, Joey Heinen accepted the Digital Production Coordinator position within the Scholarly Communication department at Northeastern University.

6. REFERENCES

[1] Hutar, J. 2013. Assessing Digital Preservation Strategies. Archives New Zealand (Wellington, New Zealand). http://ica2012.ica.org/files/pdf/Full%20papers%20upload/ica12Final00155.pdf

[2] National Digital Stewardship Residency (NDSR-Boston). 2015. http://projects.iq.harvard.edu/ndsr_boston

[3] Harvard Library. "Overview: DRS & Delivery Systems". http://hul.harvard.edu/ois/systems/drs/

[4] PREMIS Editorial Committee. 2011. Introduction and Supporting Materials from PREMIS Data Dictionary for Preservation Metadata, version 2.1. http://www.loc.gov/standards/premis/v2/premis-report-2-1.pdf

[5] Gattuso, J. and McKinney, P. 2014. Converting WordStar to HTML4. In *iPres 2014* proceedings. Archives New Zealand (Wellington, New Zealand). http://ndha-wiki.natlib.govt.nz/assets/NDHA/Publications/2014/WordStar-ipres2014-4.pdf

[6] Schaller, M. and Schlarb, S. 2014. SCAPE: Large Scale *Research and Development Department* (Vienna, Austria). https://onbresearch.wordpress.com/2014/06/16/scape-large-scale-image-migration/

[7] Bajcsy, P., Kooper, R., Marini, L., McHenry, K. and Ondrejcek, M. 2010. A Framework for Understanding File Format Conversions. In proceedings for *Roadmap for Digital Preservation Interoperability Framework Workshop*. University of Illinois at Urbana-Champaign (UIUC) (Champaign, IL). http://dl.acm.org/citation.cfm?doid=2039274.2039284

[8] Becker, C., Kulovits, H. and Rauber, A. 2010. Trustworthy Preservation Planning with Plato. In *European Research Consortium for Informatics and Mathematics, Is. 80*. pp. 24-25. Technical University Vienna/AARIT (Vienna, Austria). http://ercim-news.ercim.eu/images/stories/EN80/EN80-web.pdf

[9] Luan, F., Nygard, M., Sindre, G., and Aalberg, T. 2011. Using a Multi-Criteria Decision Making Approach to Evaluate Format Migration Solutions. In conference proceedings for *MEDES '11*. San Francisco, California. http://dl.acm.org/citation.cfm?doid=2077489.2077498

[10] Burns, P., Madden, T., Girogianni, E. and Williams, D. 2005. Migration of Photo CD Image Files. In conference proceedings for *IS&T: The Society for Imaging Science and Technology*. East Kodak Company (Rochester, New York). http://losburns.com/imaging/pbpubs/43Arch05Burns.pdf

[11] Jack, K. (2005). Color Spaces. In *Video Demystified: A Handbook for the Digital Engineer* (4th Edition, pp. 15-34). Elsevier. http://www.compression.ru/download/articles/color_space/ch03.pdf

[12] McGuffog, S. 2015. pcdMagic User Manual. https://sites.google.com/site/pcdmagicsite/

[13] Felix, T. 2009. Software the *Really* Supports Kodak Photo CD. http://tedfelix.com/PhotoCD/PCDSoftware.html

[14] Eastman Kodak Company. 1992. Image Pac Compression and JPEG compression: What's the Difference?

http://www.kodak.com/digitalImaging/samples/imagepacVsJPEG.shtml

[15] Kodak Professional. 2000. Using the ProPhoto RGB Profile in Adobe Photoshop v5.0. http://scarse.sourceforge.net/docs/kodak/ProPhoto-PS.pdf

[16] Hill, B., Roger, T. and Vorhagen, F.W. 1997. "Comparative analysis of the quantization of color spaces on the basis of the CIELAB color-difference-formula. In *ACM Transactions on Graphics, Vol. 16 Is. 2* pp. 109-154. http://portal.acm.org/citation.cfm?doid=248210.248212

One Core Preservation System for All your Data. No Exceptions!

Marco Klindt
Zuse Institute Berlin (ZIB)
Takustr. 7, 14195 Berlin
Germany
klindt@zib.de

Kilian Amrhein
Zuse Institute Berlin (ZIB)
Takustr. 7, 14195 Berlin
Germany
amrhein@zib.de

ABSTRACT

In this paper, we describe an OAIS aligned data model and architectural design that enables us to archive digital information with a single core preservation workflow. The data model allows for normalization of metadata from widely varied domains to ingest and manage the submitted information utilizing only one generalized toolchain and be able to create access platforms that are tailored to designated data consumer communities. The design of the preservation system is not dependent on its components to continue to exist over its lifetime, as we anticipate changes both of technology and environment. The initial implementation depends mainly on the open-source tools Archivematica, Fedora/Islandora, and iRODS.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges; Preservation strategies and workflows; Innovative practice.

Keywords

Contexts of preservation, data model for management and access, preservation strategies, infrastructure, Archivematica, iRODS, Fedora/Islandora.

1. INTRODUCTION

Digital data and information is not only ubiquitous but also more and more the foundation for research, education, and dissemination of cultural heritage. A lot of effort is put towards digitization of cultural artefacts in galleries, libraries, archives, and museums (GLAM), but the lack of institutional resources to keep these substantial investments not only safe but usable for future generations raises the demand for preservation services significantly. Even though the core business of these institutions is to preserve their physical holdings and collections, some of them are unable to provide and maintain adequate custody/stewardship of *digital* objects.

For that reason, we identified a need for a long-term digital preservation archival information system (DPS) that supports cultural heritage and research institutions. They can use it as a service that offers the possibility to safeguard their digital artefacts without putting many resources towards implementing the rather complex requirements for best-practice digital preservation themselves. The preservation system we describe is intended to be trustworthy in the sense that it is transparent in its functionality, documentation, and policy and is also aligned to the Open Archival Information System (OAIS) functional model as standardized in ISO 14721:2012[9]. Clients of black box preservation systems may experience difficulties in risk-assessing the underlying processes and services, and scheming exit strategies. For a more thorough explanation of concepts and terminology refer to the OAIS magenta book[7] or Lavoie's Introductory Guide[4].

This paper will first establish the background and requirements, then give a description of and reasoning for the architectural design, followed by details about the implementation and tools chosen, discussing the implementation details in regard to the design and conclude with future work and discussion.

1.1 Background

The Zuse Institute Berlin (ZIB) is a research institute for applied mathematics and computer science and operates a regional super computing center that requires facilities to store more than five petabytes of data on disk and nearly a hundred petabytes on tape. Data needed for and generated by super computer runs are expensive and therefore great effort is made to ensure reliable data retention. Our working group utilizes the available infrastructure to design and build an archival information system.

The DPS should offer deposit, curation, and preservation services for any kind of data that data producers (either in-house or external) would want to keep safe-guarded. A data-agnostic view should enable us to utilize a single core workflow system for digital objects from various domains. The system combines existing free and open-source components including Archivematica[12], iRODS[3], and Islandora[5] to vertically integrate the existing infrastructure.

We also provide a generic access layer to the submitted data for administrative and preservation watch purposes subject to access control mechanisms. The architecture is designed to support future access mechanisms for external users.

Data from cultural heritage institutions is perhaps the most universally approachable data, because it is suitable for reuse not only in scientific work but is also of wide interest for educational or creative purposes. In contrast to research data in the natural and life sciences, where data often comprises unique numeric data sets only usable in very specialized domains, and the humanities, which often is concerned with textual data, the standardization of metadata descriptions for cultural heritage objects is paramount for discovery, comparability and reuse in the semantic web. We try to accommodate data deposits from all LAM institutions by accepting *Metadata Object Description Schema*[8] (MODS) for libraries, *Encoded Archival Description*[11] (EAD) for deposits from archives, and *Lightweight Information Describing Objects*[2] (LIDO) for museum object descriptions as preferred formats.

1.2 Requirements

In addition to the special cases of cultural heritage the system should be able to handle data from all fields in research and education and to help them to maintain the viability of the vast corpus of digital materials either already amassed or in the process of being generated.

The main requirement is to maintain deposited information as self-contained and self-describing *archival information packages* (AIP) using the preservation metadata dictionary as described by PREMIS[10]. In an ideal world AIPs should not depend on the existence of any component of DPS itself and therefore enable the exchange of AIPs within a collaborative federation of OAISs as described in [13]. Unfortunately this is not the case; interoperability in terms of AIP exchange between systems is a complex problem which remains to be solved.

The system should furthermore use or adapt existing and available open source tools and open standards and by doing so benefit from community best-practices and advancements. Implementing a DPS from scratch and keeping up with research is not feasible.

1.3 Review of existing solutions

Existing, commercially available preservation systems do not fully meet our workflow requirements due to lack of access to the source code and availability of publicly accessible documentation, or mainly depend on cloud infrastructure. Opensource OAIS aligned systems seemed too complex to easily change components without losing functionality. Integrating different components into one system offers the opportunity for clear responsibilities, audit, and documentation and therefore trustworthiness.

2. ARCHITECTURAL DESIGN

Because of our requirements to be aligned to OAIS and have self-contained AIPs for interoperability, we designed the architecture to be a single, modular core pipeline of existing tools linked up by in-house developed data conduits.

The system is therefore composed of loosely coupled components with strictly defined responsibilities. As we anticipate the software components to become obsolete during the lifetime of the system, this modularity enables us to find or develop tools to substitute waning functionalities.

Loose coupling also means that redundancies with regard to data and metadata are necessary to achieve the goal of independent, functional exchangeable modules. The self-contained, homogenous information packages do also support a more streamlined, automatical migration of archived content into other DPSs as an exit strategy.

2.1 Aiming at moving targets

The whole system design and implementation is regarded as a living system that has to be cared for and adapted to a changing environment and requirements. We try to achieve this goal by utilizing a single core preservation workflow for all information packages to reduce complexity and increase sustainability. The DPS as a whole consists of the stages *preingest* (see 2.2), *ingest* (2.3), *management* (2.5), and *access* (2.6). Only the first and last stage (deposit and access) will be customized to meet the requirements of different types of data, the core stages (ingest and management) will process the data from all producers the same way. An overview of the system architecture is shown in figure 1.

A consistent and well-defined data model is a fundamental prerequisite that allows for not only treating all submitted data with a single workflow but also, more importantly, to supersede tools in the future. Our data model distinguishes between *preservation description information* (PDI) and *descriptive information* (DI) as described by the OAIS. The PDI in our case also contains descriptive and administrative information about the data producer such as identifiers, contract numbers, contacts, and more. See figure 2 for an overview of the data model.

2.2 Preingest

The ingest functional entity described in the OAIS model is split into two phases within our architecture: deposit and ingest. The deposit phase covers the preingest process, i.e. the preparation and transfer of data as a submission information package (SIP) into a quarantine staging area. The ingest phase covers the preparation of the archival information package (AIP).

The DPS does require data to be organized and described in a certain way in order to treat the deposit's content agnosticly independent from origin and purpose. The deposit workflow ensures completeness of administrative metadata and transforms the data formats into the data model expected by the ingest process. The original data is preserved to maintain authenticity.

Information packages preserve only the included information. To preserve the ability to independently understand the contained information requires that the data can be rendered as information by the consumers. Therefore the producer has the responsibility to ensure that the data is suitable for re-use by a designated community. The responsibility of the DPS is to maintain that renderability of the data. With research data sets this often conflicts with the need for widely supported representation information, i.e. file formats, that are easily rendered with standard soft-

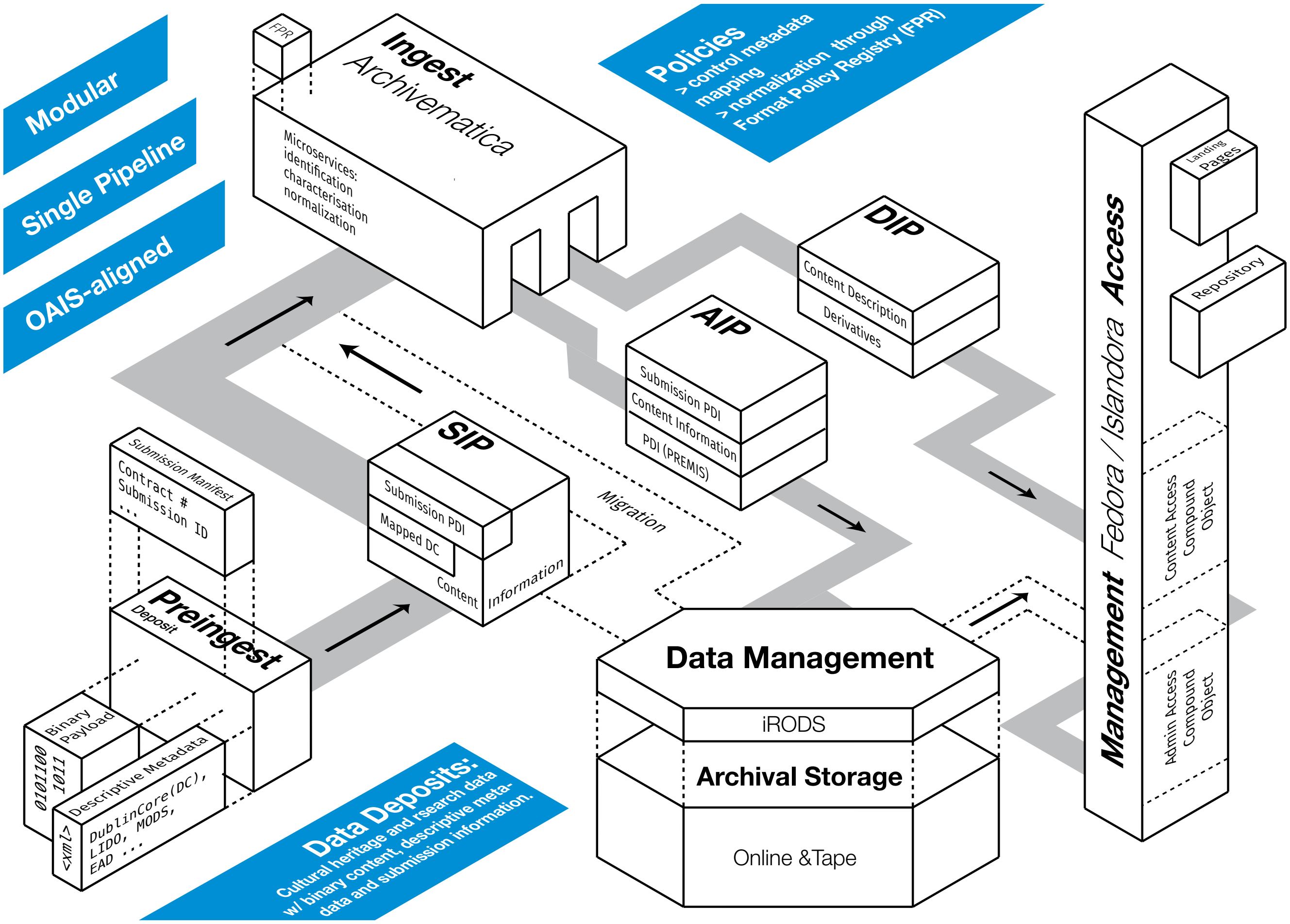


Figure 1: Digital Preservation System Architecture Overview

ware. To support the preservation of any data regardless of format, our preservation system differentiates only two levels of preservation: passive and active. Data perceived to be at the passive preservation level will be preserved at the bit-level in addition to retaining structure and metadata, and the DPS promises a best effort to describe the contained data. Data perceived to be at the active preservation level requires the best effort of a depositor to comply with policy-published archival formats and the DPS therefore promises to ensure the renderability through migration.

The system will accept any digital material for deposit but rejects any submission for ingestion that does not satisfy the submission agreements. Preservation activities will differ based on the assessment of a preservation level on ingest. The perceived level of preservation is however not a static one, but is an outcome of re-examination of the supplied or extracted technical metadata. It can change from passive to active preservation as a result of an updated format policy following actionable observations during preservation watch.

Context information, which might be useful for understanding the deposited material in the future but is not part of the information object and therefore does not have to be considered for preservation actions, can be declared to be submission documentation. This data will be captured in the AIP but is not included in the DIP.

During the deposit phase all necessary metadata is gathered that is needed for managing and accessing the data within the archive.

2.2.1 *Data Deposit Registration*

Prior to depositing the content information itself, the producer must initiate a data deposit session by requesting a submission identifier through a web portal. The submission identifier will act as the reference for the data to be deposited and will be used to attach the transfer data to the deposit agreement negotiated between the producer and our archive. The deposit agreement itself comprises a legal contract for transfer of custody including responsibilities of the producer and the DPS, and a technical policy (the submission agreement) describing workflows, procedures, and actions based on the types of data objects. Upon registration the producer will have the choice either to input mandatory and optional preservation description information (PDI) via a web form or by selecting to include said information as a submission manifest with the data transfer. The PDI is necessary for managing the data by providing information about relation to contracts and submission agreements.

Descriptive information (DI) in Qualified Dublin Core can be entered or included as CSV[1] or as XML for Qualified Dublin Core in the data transfer. If the descriptive metadata is included as either *Encoded Archival Description* (EAD) for deposits from archives, *Lightweight Information Describing Objects* (LIDO) for museums, or MODS for libraries, the necessary fields will be automatically extracted and mapped to DC during the restructuring stage.

The submission session can be interrupted and resumed at any time to allow for thorough preparation of data objects and content information. If for any reason the producer decides to abandon the session altogether he can also terminate it and discard entered and uploaded data.

2.2.2 Data Transfer

The data objects can now be uploaded to the staging area which is provided alongside the submission identifier. The data can either be uploaded via a web browser, transferred to the staging area by other network protocols or via sneakernet on external hard disk drives, USB thumbdrives or optical media. To ensure integrity and completeness during transfer the data must reside either in ZIP archive containers, be put into a BagIt structure or be referenced in a METS files with checksum information in the file section.

2.2.3 Submitting and Compliance Testing

In order to conclude the submission session, the producer must initiate a compliance test by clicking the submit button. This checks for completeness of the required preservation description information and the presence of descriptive information. If ambiguous data is detected during the automatic metadata extraction, the submission is considered not to be compliant.

Furthermore the integrity of all data objects is checked by testing the zip containers, validating the BagIt integrity or checking the uploaded files against the file section of the METS files.

If the submitted data deposit successfully passes these tests, the data is accepted and transferred to the restructuring step.

2.2.4 Restructuring

The quarantined data is restructured into either a single or multiple submission packages. Some producers with poor data management facilities or legacy applications choose to bulk export data sets and upload corresponding content information and rely on the archive to bundle the appropriate files and metadata into information packages. The rules for breaking up bulk deposits into multiple information packages are specified in the submission agreement.

The descriptive information, either as entered in the registration process or extracted from metadata files, and the preservation description information are transformed into a format suitable for ingest. The original metadata files are treated as data objects and bundled with the remaining data objects. METS files that were used for transfer and do not contain descriptive information are put into the submission documentation area.

After restructuring all SIPs are of equal structure and ready to be ingested through a single ingest workflow.

2.3 Ingest

The ingest phase creates an identifier for the AIP and assigns identifiers to all data objects for reference within the AIP. The ingest workflow identifies common file formats and extracts necessary technical information. Non-archival file formats are normalized if the delivered content is identified to be in a set of known formats for which format policies exist. Only content already in a set of archival formats or conforming normalized versions can be actively preserved. Content not identified in ingest will only be preserved passively, i.e. at the bit-level. This enables the archive managers to easily identify the need for migration preservation actions and their planning.

PDI including rights statements, DI, logical and physical structure of the SIP, fixity information, PREMIS events of identification and normalization is captured into a single METS file that will be the authoritive source of information about the AIP for managing the archive. All files are compiled into a BagIt structure that is saved as a single archive file to allow for easy transfer within the data management layer.

Access or dissemination copies of either the normalized data objects or derivatives are also created during ingest and transferred along with DI and PDI to the combined management and access repository.

A submission report will be sent to the contact person (producer) by email. The data deposited is now under stewardship of the archive.

2.4 Archival Storage

The AIPs are transferred to archival storage by a data management middleware layer. The middleware abstracts from the physical resources and is responsible for not only storing AIPs as multiple replications to on-site and potential off-site locations but also to retrieve the physical AIPs independent of residence. A subset of the PDI is attached as administrative metadata to the AIPs such as the type of information package, the identifier, submission and contract identifiers and fixity information. Aside from using this metadata in the management of the archive, it can also be used as the database for generating reports on storage usage or item count reports.

2.5 Management

The higher-level data management operations (see 3.4) are not based on information stored in the data management middleware, but based on the PDI and DI stored in a repository alongside the dissemination copies of the content information. There are two main roles for accessing the information contained in the DPS: data managers and data consumers. Data managers are entrusted with keeping data usable. Data consumers want to search, discover, and retrieve information. To address the different needs of these

[1] Comma-separated values. See IETF RFC 4180.

roles the repository provides two different views to the same AIP realized as separate entities within the repository object store: an administrative entity providing the PDI and a descriptive entity providing the DI. Management activities like monitoring, reporting or performing preservation actions often require selection of AIPs through information contained in the PDI, which is accessible through the administrative view. The DI entity is responsible for access by data consumers.

2.5.1 Preservation actions: migration

Following a change of policy regarding a certain file format, the PREMIS records are checked for occurrences of that format and the corresponding AIPs are selected for migration. The change in format policy affects the normalization step in the ingest workflow for all subsequent submissions. Migration is performed by re-ingesting selected AIPs as SIPs into the ingest workflow preserving the identifiers and amending the PREMIS event trail.

2.6 Access

The access repository as mentioned contains not only representational copies of the data objects but also the corresponding DI and PDI. Access to information in the PDI and the retrieval of AIPs residing in archival retention is restricted to data managers. Data consumers on the other hand have different requirements: the finding of information inside the archive based on the descriptive information provided in DC and a representation of the content information suitable for their needs. The access to DIPs is restricted based on the access rights information from the PDI. To provide different designated communities or other users a higher level of precision and recall for retrieval, subsets of DIPs can be transferred to access systems or repositories that provide a more specialized integration and understanding of the original descriptive metadata schema instead of generic Dublin Core.

If the access to a data package is assigned an open license, a Digital Object Identifier (DOI) can be provided for persistent access to research data sets at the DIP level.

3. IMPLEMENTATION

We chose to make use of existing and freely available tools as much as is feasible to achieve the architectural design goals and to keep in-code customization of existing tools and inhouse development to a minimum. The transfer and the access stages are obviously the most challenging because of our goal to accept data deposit from various domains. After restructuring the delivered information packages and mapping of metadata, a *single* pipeline is used for preservation and management actions.

The normalized deposits are processed by a single Archivematica pipeline for AIP generation, handed off to an iRODS data management grid for archival storage and transfer, and the PDI and a subset of DI and CI (i.e. derivatives where appropriate) are ingested in a Fedora object storage, from where they are accessible through an Islandora front end for both management, discovery, and retrieval functionality.

The implementation is guided by the overall architectural design and adapts to lessons learned during the development. Some stages are not completely functional yet.

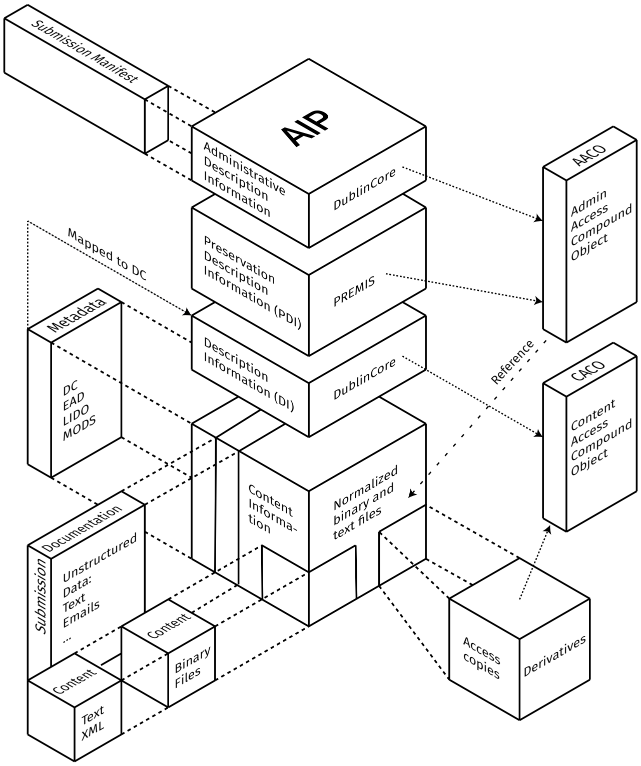


Figure 2: Data Model and AIP

3.1 Preingest

While the self-deposit of data will be available in the future via a web portal as described above, most submitted data arrives currently by external media or secure file transfer (sftp) or copy (scp). The integrity is checked by verifying a bag structure or zip archive created by the producer. The data resides in a quarantine storage area (and temporary backup) and is available for the data extraction and restructuring tools.

The administrative PDI metadata (an excerpt is shown in table 1) for the submission is stored in a submission manifest file in YAML[2], which is human and machine readable. Based on the field `MetadataFormat`, a customized python script is selected to extract descriptive DI metadata for content discovery from `MetadataFile`. Supported metadata schemas are EAD, LIDO, MODS, and qualified DC-XML. Extraction of DC from a METS container will be supported in the future. The scripts also check for the completeness of the payload that is referenced in the metadata file.

After having successfully gathered the PDI and the DI, all files will be sorted into either the object or the submission documentation directories of a single or multiple SIP depending on rules in the submission agreement. The script

[2] YAML is a human-friendly data serialization standard. See `http://yaml.org/`.

Field name	Dublin Core Mapping
SubmittingOrganization	dcterms:rightsholder
OrganizationIdentifier	dcterms:publisher
ContractNummer	dcterms:identifier
Contact	dcterms:creator
...	
AccessRights	dcterms:accessrights
License	dcterms:license
...	
Metadatafile	not mapped
Metadataformat	not mapped

Table 1: Submission manifest and mapping

also writes metadata to special files that are processed by the ingest workflow. The subsequent stages will operate on these now equally structured SIPs.

3.2 Ingest

The ingest phase will transform the submitted data and metadata, extracted technical metadata and the documentation of the actions taken into an AIP. After extensive market research and testing we found that Archivematica is a viable tool that meets our requirements quite well. Archivematica is an open-source application that combines various open-source tools into a distributed workflow pipeline to process digital objects into AIPs following the OAIS functional model.

3.2.1 Archivematica

The ingest phase will be executed by a *single* Archivematica pipeline. The processing workflow can be controlled with a web interface or through REST calls. Creating entirely new workflows or modifying the existing is cumbersome at the moment because it is stored in a database with heavy use of referencing. The existing workflow contains workflow decision points that can be influenced by presets, and these will be applied to all deposits. The preconfigured choices can be overridden by embedding a preset file into the transfer data directory. This allows, for example, to store AIPs in distinct storage locations if negotiated in a submission agreement.

Archivematica generates a METS structure to capture the references to files and file structure of the ingest, to attach metadata to individual files, and to document rights and executed preservation actions and their result in PREMIS events. Metadata is attached to files in the submission by creating a metadata CSV file that Archivematica then inserts into the respective descriptive metadata section (DMDSec) in the METS. We wanted to keep changes to the codebase of Archivematica to a minimum but also have control over metadata that end up in the generated METS to support the PDI from the submission manifest. The METS standard schema does not support the description of the METS file itself, so we use the METS generation script in Archivematica to attach the administrative data to the object directory level as a convention for the metadata to survive the ingest. As we have control over the ordering of the DMDSecs we also use the second DMDSec to store the DI about the submission. The structure of the resulting AIP is shown in figure 2.

Archivematica also generates DIPs that are not handed off to archival storage but are used to ingest data into the access and management repository (see 3.4.1).

3.3 Archival Storage

For archival storage we use the on-site data storage facilities available at ZIB. Persistent storage consists of a hierarchical Storage Archive Manager (SAM) that augments an online file system transparently with nearline redundant tape storage. ZFS is used for the online filesystem that is designed to prevent data corruption caused by bit-rot. Nearline access to tape storage means offline data is available in less than 30 seconds on average. The data is stored redundantly on two StorageTek 8500 tape libraries with currently installed tape capacity of around 100 petabyte, of which nearly 400 terabyte (800 terabyte with redundancy) are reserved for our archival system. The libraries have no physical connection to prevent tapes from being destroyed by tape recorder malfunctions. They also use automatic fixity checks and error correction to ensure data integrity. The tape libraries are installed in a special vault that is waterproof, can withstand an outside fire for around 10 hours, and has an additional CO_2-fire extinguisher system.

Archivematica supports different storage backends through the use of separate storage service application that abstract various services like local and NFS file systems, LOCKSS, Duraspace, and others. As we use iRODS (integrated Rule-Oriented Data System) to store and replicate digital objects, we extended the Archivematica storage service to expose an iRODS storage space which not only stores AIPs but also attaches administrative metadata to help with discovery and retrieval.

3.3.1 iRODS

iRODS is a distributed data-management system for creating data grids and persistent archives. It provides access to data objects organized in collection trees called zones with granular access control. Data in a zone can be accessed by authenticated users regardless of where the data is physically stored. Integrated rules manage replication to physical storage resources transparently to the user and can also act on user-supplied metadata attached to the data objects. Such replication is also possible to remote, off-site storage for geographical redundancy. The iRODS grid supports integrating storage resources and user bases of different organizations and thus can be used for federated archiving.

iRODS tiered resources are responsible for replicating AIPs to the SAM and back into online storage. iRODS maintains checksums for all AIPs that are used for fixity checks if AIPs reside online. Online storage is more expensive than tape so rules are implemented to trim redundant copies of AIPs if a threshold of disk usage is reached.

Although federated data replication is also possible with LOCKSS, by using iRODS we maintain more control over data movement, residence and replication.

3.4 Management

High-level management of the data in our DPS consists of monitoring data integrity, triggering preservation actions, and providing access to AIPs and DIPs. Preservation actions that migrate file formats are not yet implemented but one of the future releases of Archivematica will add a feature that will allow us to re-ingest AIPs back through the Archivematica pipeline. This allows for changing metadata in already ingested AIPs without changing the identifiers including amendment of the PREMIS trail. Extending the feature to re-normalize file formats for which the format policy has changed in Archivematica is planned for the future.

3.4.1 Fedora/Islandora

We ingest the DIPs generated by Archivematica into a Fedora Object Store and use Islandora as a front end to the repository for management actions. The DIP contains the same information in METS as the stored AIP including the PREMIS data and derivative representation (access copies) of the binary payload (where appropriate), and a transformation has been defined and has been carried out by the ingest stage.

We represent a single DIP with *two* Fedora compound objects: one for administrative and management purposes and another for content discovery and access purposes. The DIP METS is parsed on ingest and transformed into multiple Fedora METS files that are ingested as multiple Fedora objects: one METS for each binary payload file or access copy, and one METS for each of the two compound objects as parents. One of these parent objects will be ingested as an *admin access compound object* (AACO) and comprises all data streams contained in the DIP. The AACO gives access to the submission manifest data as main descriptive Dublin Core and refers to the PREMIS data and payload derivatives. It is used for administrative tasks involving contracts and deposits. Additionally the payload derivatives and the DI, i.e. the description of the digital object, are accessible through another object called *content access compound object* (CACO). This is used to discover and retrieve objects based on the mapped, generic DC metadata of the datasets. These two different ”views” of a DIP try to separate administrative tasks like report generation of stored file formats or calculating the amount of data stored for each contract or year, and finding objects by their actual content while keeping the datastreams in a unified repository. The AIP is referenced through the Islandora front end by the identifier of the AIP and can be retrieved from the storage layer through the AACO view. The iRODS integration with Islandora and the data model is described in more detail in [1].

The DIPs managed by Fedora are stored in a filesystem that is backed up separately from archival storage. They can, however, be re-generated any time from the AIPs.

3.5 Access

The repository is currently accessible only for internal administrative users. Access to the AIPs for data producers is realized with management actions by the administrative staff to stage data in a location where it can be picked up; we provide no self service at the moment. With changing requirements of our clients we might have to implement an access repository for bulk self-service AIP exports in the future.

The descriptive information exposed through the repository is limited due to its DC-only design and therefore not adaptable to the different metadata descriptions from the various domains, we plan to support, and is therefore not suitable for discovery and reuse of the data within those domains. This is a consequence of the need for normalization of descriptive metadata for utilizing a single preservation system.

We do provide DIP presentation access for selected data sets not from within the internal access repository but by generating landing pages or handing data off to external content management systems (CMS) or repositories.

3.5.1 Landingpages

For some clients who have no means of providing access to their data themselves, we offer a service for generating customized, static landingpages for each AIP as a low-maintenance way to present them on the Web. The customization includes converting the used metadata schema to HTML templates and populating them with metadata and binary data from the DIPs. Discovery can be provided by generating a digital object identifier (DOI) for reference in publications or uploading metadata to specialized search portals, e.g. Europeana[3] or the Deutsche Digitale Bibliothek[4] (DDB, German Digital Library) for cultural heritage data. As the landingpages are static, they can easily be migrated to other hosting services as they are independent from our infrastructure.

4. FUTURE WORK

New data producers who want to use our infrastructure might require different protocols for data deposits. We will investigate the suitability of SWORDv2 (Simple Web-service Offering Repository Deposit)[6] or OAI-PMH[5] for data from institutional repositories or the S3[6] data cloud protocol for large research data sets.

Collaboration with other archives will be tested by taking over their AIPs by transforming their structure in our deposit stage and ingesting them into our pipeline. Other archives can also ingest our AIPs in their archives because the information packages are self-contained and do not depend on data residing in databases. Other archives could also choose to use only the ingest stages or utilize our archival storage through the federation and replication features in iRODS.

4.1 Object repositories

For clients who require the data to be stored in and accessed through a data repository, we may potentially offer Islandora instances that can be customized to their designated user communities and corporate designs. The CACO and its relevant child objects would then also be transferred to these repositories after ingest.

For clients with existing CMS or repositories but without the infrastructure or resources for digital preservation we

[3] http://www.europeana.eu/portal/

[4] http://www.deutsche-digitale-bibliothek.de/

[5] Open Archives Initiative Protocol for Metadata Harvesting

[6] Amazon Simple Storage Service API, a protocol implemented in various technology stacks.

plan to hand-off the DIPs into their CMS or repositories so that they have an overview about the deposited data. These repositories could be used in the future to generate AIP delivery orders.

The original (not mapped) metadata description schema contained in the AIP could also be used for the development of sophisticated platforms to explore and discover the data because it takes advantage of the inherent complex data models they are based upon.

5. DISCUSSION AND CONCLUSION

Our data model for metadata supports digital long-term preservation within a single core workflow for ingest and management activities and allows for consistent description of widely varied content and a clear separation of PDI and DI. The resulting AIPs accommodate both submission requirements from multiple data producers as well as accommodate discovery opportunities for data consumers in addition to the information needed for administration, access control, preservation watch (using the PREMIS information), reporting, and billing for the management entity.

The core idea of AIP construction is to treat the AIPs as well-documented, atomic information objects that contain the full intellectual information about the preserved objects without external references. This obviously permits a simple exit strategy in case the DPS ceases to exist in the future.

The described model, design, and implementation may not be suitable for everyone. However, we hope that it enables us to offer preservation services to a whole range of different data producers because it reduces complexity of the infrastructure and therefore helps manageability and sustainability of the whole system. The modular architecture allows us to substitute software building blocks as reaction to technical issues related to software obsolescence. At the same time it deals with the intrinsic complexity and variety of data and contained information that has to be preserved in order not to deprive future users of possibilities and opportunities.

6. ACKNOWLEDGMENTS

This work has been supported by the Senate Chancellery Cultural Affairs of the State Berlin. The authors would like to thank Tim Hasler, Wolfgang Peters-Kottig, and Elias Oltmanns for valuable input and discussions.

7. REFERENCES

[1] K. Amrhein and M. Klindt. Islandora as an access system for iRODS managed information packages. Presented at the 10th International Conference on Open Repositories, 2015.

[2] E. Coburn, R. Light, G. McKenna, R. Stein, and A. Vitzthun. LIDO - Lightweight Information Describing Objects Version 1.0. Technical report, ICOM-CIDOC Working Group Data Harvesting and Interchange, November 2010.

[3] B. Fortner, S. Ahalt, J. Coposky, K. Fecho, A. Krishnamurthy, R. Moore, A. Rajasekar, C. Schmitt, and W. Schroeder. Control Your Data: iRODS, the integrated Rule-Oriented Data System. White paper, RENCI, University of North Carolina at Chapel Hill, 2014.

[4] B. Lavoie. The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition). *DPC Technology Watch Series*, Oct 2014.

[5] M. A. Leggott. Islandora: a Drupal/Fedora Repository System. In *4th International Conference on Open Repositories*, May 2009.

[6] S. Lewis, P. de Castro, and R. Jones. SWORD: Facilitating Deposit Scenarios. *D-Lib Magazine*, 18, January/February 2012.

[7] Reference Model for an Open Archival Information System (OAIS). Hosted at public.ccsds.org/publications/archive/650x0m2.pdf, 2012.

[8] Metadata Object Description Schema (MODS). Hosted at http://www.loc.gov/standards/mods/. Retrieved September 2014.

[9] ISO 14721:2012: Space Data and Information Transfer Systems – Open Archival Information System (OAIS) - Reference. Retrievable via http://www.iso.org/iso/catalogue_ics, 2012.

[10] Preservation Metadata: Implementation Strategies (PREMIS). Hosted at http://www.loc.gov/standards/premis/. Retrieved March 2014.

[11] B. Stockting. Time to Settle Down? EAD Encoding Principles in the Access to Archives Programme (A2a) and the Research Libraries Group's Best Practice Guidelines. *Journal of Archival Organization*, 2(3):7–23, July 2004.

[12] P. Van Garderen. Archivematica: Using micro-services and open-source software to deliver a comprehensive digital curation solution. In *Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria*, 2010.

[13] E. Zierau and N. Y. Mcgovern. Supporting Analysis and Audit of Collaborative OAIS's by use of an Outer OAIS – Inner OAIS (OO-IO) Model. In *Proceedings of the 11th International Conference on Digital Preservation, Melbourne, Australia*, October 2014.

Applying Translational Principles to Data Science Curriculum Development

Liz Lyon
School of Information Sciences
University of Pittsburgh
+1 412 624 9436
elyon@pitt.edu

Eleanor Mattern
School of Information Sciences
University Library System
University of Pittsburgh
+1 412 648 5908
emm100@pitt.edu

Amelia Acker
School of Information Sciences
University of Pittsburgh
+1 412 624 4939
aacker@pitt.edu

Alison Langmead
Dietrich School of Arts & Sciences
School of Information Sciences
University of Pittsburgh
+1 412 648 2407
adl40@pitt.edu

ABSTRACT

This paper reports on a curriculum mapping study that examined job descriptions and advertisements for three data curation focused positions: Data Librarian, Data Steward / Curator, and Data Archivist. We present a transferable methodological approach for curriculum development and the findings from our evaluation of employer requirements for these positions. This paper presents "model pathways" for these data curation roles and reflects on opportunities for iSchools to adopt translational data science principles to frame and extend their curriculum to prepare their students for data-driven career opportunities.

General Terms

Training and education.

Keywords

Curriculum development, translational data science, research data curation, iSchools.

1. INTRODUCTION AND CONTEXT

The growing focus on data science, research data management services and associated data curation and preservation strategies represents evidence of the increasing operational impact of the data deluge, the need for data infrastructure development and a realization that significant workforce capacity and capability challenges are emerging. Employers in all sectors are seeking graduates to fill a diverse mix of data-related roles, characterized by a broad range of data literacy skills and competencies, combined with disciplinary knowledge and practical experience. In this complex landscape, Information Schools (iSchools) are reviewing curriculum requirements and developing new data-centric courses to build capacity in the workplace and to support data-driven careers in the 21st century.

This paper will present the outcomes of a curriculum mapping exercise, which has built on translational principles (i.e. "*translating research into practice"*) [1] and has recognized the distinctiveness of different data science roles. The methodology is transferable, with a particular emphasis in this paper on career options in data curation and preservation; we highlight opportunities for innovative course offerings and the development of new educational collaborations and partnerships.

2. RESEARCH QUESTIONS

The study addresses three specific areas of interest:

1. What are the skills, competencies, knowledge, experiences and education required for the distinct data science roles?
2. How do these data science role requirements map to current curriculum topics and course offerings?
3. What opportunities emerge for new collaborations and partnerships in developing the data science curriculum?

3. LITERATURE REVIEW

The importance of data for libraries was recognized as early as 2006 [2] [3] and data librarianship was identified as a "gap in the market" in 2008 [4]. The roles and responsibilities of librarians and data were reviewed [5] [6] [7] and the need to re-skill subject and liaison librarians has been described [8], [9]. Surveys of RDM activities in libraries have been published [10], [11] which demonstrate a gradual ramping-up of infrastructure and service delivery.

Two recent reports from the UK have highlighted the requirement for "*a skilled workforce and data-confident citizens*" [12] and "*a severe shortage of UK data talent*" [13]. Whilst these reports have primarily focused on data analytics, earlier reports have described "*a dearth of skilled (data) practitioners*" [14] and "*the current paucity of data scientists*" [15], recognizing contrasting roles and responsibilities. Marketplace analysis of trends data from Indeed.com, demonstrates a steady growth in data-related positions [16], reinforcing the demand : supply ratio imbalance. Examining the range of nomenclature for describing these positions, "a *need to disambiguate and develop definitions for professional roles*" [17] has been recognized. Four discreet data roles were identified by Swan and Brown [15]. In this paper we will draw on the family of (six) data scientist roles described by Lyon and Brenner [18]: data librarian (University or Research Institute), data archivist (National Archive), data steward (Data Center), data analyst (Corporate sector), data engineer (IT Company) and data journalist (News & Media). Typical employment locations are indicated in brackets; this is a key perspective since the tangible requirements of real-world settings, are positioned to be primary drivers in curriculum development. The current study examines the first three roles in depth and will draw out perceived commonalities and differences.

The range of data-related roles is reflected in the breadth and depth of the data curriculum, since its scope should support the specific requirements of each role. The position, function and value of iSchools in developing data workforce capability and building capacity has been noted [18]. Data-related graduate programs, certificates and courses are already provided by some institutions [19] [20]. Two new graduate courses (Research Data Management and Research Data Infrastructures) have been designed and delivered at the University of Pittsburgh iSchool, alongside data mining, data analytics and data visualization classes and an Advanced Certificate in Big Data. There are also a range of data management training programs positioned towards up-skilling existing library and information professionals e.g. RDMRose [21], immersiveInformatics [22], MANTRA DIY Kit [23]. A recent review of digital curation education and training, notes the development of a Research Data Management MOOC (Massively Open Online Course) by UNC-Chapel Hill in 2015 [24].

There is a growing body of work addressing the core skills, competencies and training requirements for digital curation and research data management. These initiatives have been variously framed as Data Information Literacy (DIL) [25] [26] and Data Management Skills Support Initiative (DaMSSI) [27]. Twelve core competencies were identified in the DIL Project: Data Processing and Analysis, Data Management and Organization, Data Preservation, Database and Data Formats, Ethics and Attribution, Data Quality and Documentation, Data Curation and Reuse, Data Conversion and Interoperability, Data Visualization and Representation, Discovery and Acquisition, Metadata and Data Description, Cultures of Practice.

The term "*translational data science*' was introduced by Lyon and Brenner [18] to describe the transition of data skills, software tools and research intelligence from the iSchool to the marketplace. This characterization is particularly relevant to the development of the broad data science curriculum, which aims to equip graduates for new community practice roles in a range of disciplines, organizations and sectors. The implications of a translational approach to the design of training programs in the clinical sciences has been described which highlights the need to understand complementary disciplines and to become immersed in (clinical) practice [28]. These requirements have resonance for data scientists of all flavors, who must combine a portfolio of data informatics skills and competencies with disciplinary knowledge and practice. An immersive approach to research data skills development has been adopted in the immersive Informatics program [22], in the clinical setting [29] and at the University of Pittsburgh [30], where students spend time in the research laboratory. A similar model has been implemented at the University of Illinois Urbana-Champaign, with an intern practicum located in a data center [31].

4. METHODOLOGY

In this small-scale study, we sought to create a transferable methodology that faculty at iSchools may use to examine and review their existing curriculum in order to ready their students for future translational (market-driven/real-life) data preservation and data curation roles.

We selected the three data-related preservation roles outlined by Lyon and Brenner [18] to provide a focus for an analysis of employer requirements: Data Librarian, Data Steward/Curator, and Data Archivist. We searched for recent job postings (published from January 2014 to April 2015) that were of "semantic equivalence" to these roles, using five job banks to locate the positions: indeed.com; *The Chronicle of Higher Education's Vitae;* ALA JobLIST; www.jobs.ac.uk; and the IASSIST Jobs Repository. These job postings are listed in Table 1. We selected these job banks based on both breadth and on tailored focus. We anticipated that indeed.com, a search engine that aggregates listings from multiple job sites, would offer breadth. The academic and library employment sites (*Vitae* and www.jobs.ac.uk) would enable us to search for positions within institutions of higher education and libraries, both of which we anticipated would be major employers for these data curation roles. Finally, IASSIST is an international organization of information professionals focused on social sciences and data; we selected the Jobs Repository because of its narrow and relevant scope.

We used keyword searching for the job banks. In the instance of IASSIST Jobs Repository, the volume of positions published in our studied time frame allowed us to visually scan the job titles for relevance. For our analysis, we aimed to locate ten job positions for each data-preservation role. We were successful in doing so for the Data Librarian and Data Steward positions but discovered a paucity of positions with "Data Archivist" as a job title. In this case, it was necessary to broaden our search and analyze positions that fell outside of the January 2014 to April 2015 time frame, and we drew upon the IASSIST Job Repository, which includes postings from 2005 to present. Even with this resource, it was necessary for us to include in our analysis one position that we read as "archival" in nature and that was located at a data archive, despite the absence of lexical equivalence (that is, the titles were different but the nature of the work similar).[1]

While we were able to access the full job descriptions for the more current positions, the IASSIST postings offered a more abbreviated job advertisement. In the cases in which a URL was available in the IASSIST job advertisements, we attempted to use the WayBack Machine to locate the original job descriptions. We found that while we could access the institutions' human resources websites, the job descriptions were not indexed.

We performed a content analysis on the job descriptions (and, when necessary, job advertisements) for our suite of job positions to identify patterns in employers' requirements for job candidates. We

[1] It should be noted that non-traditional and traditional archive positions are seeing an explosion of new names and classifications, including "digital asset managers," "digital content specialists," "digital services technician," "supervisory IT specialists." In each of these jobs people are responsible for preserving, describing and providing access to data sets at different scales. Forthcoming work by Acker will highlight these changes.

developed a coding scheme that examined five categories:

- *Education:* Academic qualifications
- *Experience*: direct, hands-on practice
- *Knowledge*: understanding of/familiarity with topics/subjects/issues
- *Skills*: ability to do an action well
- *Competencies*: proficiency with specific tools/technologies/programming languages.

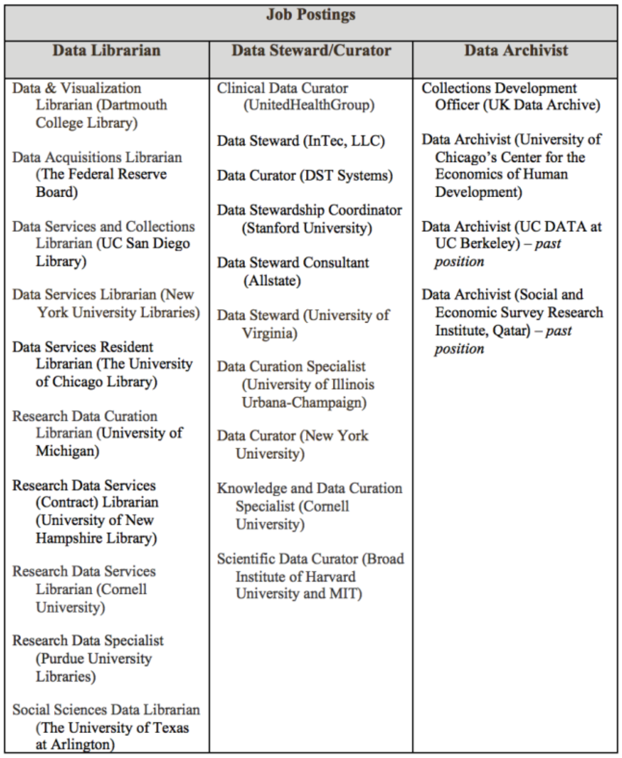
For each of the three roles studied in this paper, we sorted requirements articulated in the located job positions within these five pre-set categories. Having done this, we looked for patterns within the requirements that cut across the positions. For example, having grouped required technological "Competencies," we drew specifications that candidates be proficient with Microsoft Access, with MySQL, and Oracle and coded these specifications as "competence with relational database systems" (see Table 3).

While we made note of whether the employers characterized the education, knowledge, experience, skills, and competencies as essential or desirable requirements, we looked for patterns in the coding irrespective of this classification. In doing so, we could assess how the curriculum would best prepare iSchool students for employer consideration for the three data-preservation roles.

We identified all requirements that appeared in at least two of the positions studied for each role and designated these as "Key Requirements". From here, we analyzed course syllabi in the University of Pittsburgh School of Information Sciences graduate Library and Information Sciences (MLIS) program to determine relevant courses offered which would support the required and desired education, experience, knowledge, skills, and competencies. In doing so, we focused on the course description, objectives, and topics outlined in the syllabus. We then identified course topic gaps and opportunities for partnerships, both internal and external to the School of Information Sciences at Pitt.

As a part of the process, we explored model pathways to the data-preservation roles of Data Librarian, Data Steward/Curator, and Data Archivist, based on the current (2014-2015) curriculum. We also identified ways in which the School of Information Sciences could enhance its preparation for students to meet the expectations of employers seeking applicants for these positions.

Table 1. Job postings examined [28]



Job Postings		
Data Librarian	**Data Steward/Curator**	**Data Archivist**
Data & Visualization Librarian (Dartmouth College Library) Data Acquisitions Librarian (The Federal Reserve Board) Data Services and Collections Librarian (UC San Diego Library) Data Services Librarian (New York University Libraries) Data Services Resident Librarian (The University of Chicago Library) Research Data Curation Librarian (University of Michigan) Research Data Services (Contract) Librarian (University of New Hampshire Library) Research Data Services Librarian (Cornell University) Research Data Specialist (Purdue University Libraries) Social Sciences Data Librarian (The University of Texas at Arlington)	Clinical Data Curator (UnitedHealthGroup) Data Steward (InTec, LLC) Data Curator (DST Systems) Data Stewardship Coordinator (Stanford University) Data Steward Consultant (Allstate) Data Steward (University of Virginia) Data Curation Specialist (University of Illinois Urbana-Champaign) Data Curator (New York University) Knowledge and Data Curation Specialist (Cornell University) Scientific Data Curator (Broad Institute of Harvard University and MIT)	Collections Development Officer (UK Data Archive) Data Archivist (University of Chicago's Center for the Economics of Human Development) Data Archivist (UC DATA at UC Berkeley) – *past position* Data Archivist (Social and Economic Survey Research Institute, Qatar) – *past position*

5. RESULTS

This section presents the analysis of employer-specified job requirements for the selected roles of Data Librarian, Data Steward/Curator, and Data Archivist. Two perspectives are drawn out: a) the common requirements across the three roles and b) the requirements unique to each role. This analysis is followed by the development of specific data-centric model pathways for each role, based on the analysis of job posting requirements. We draw upon the current (2014-2015 academic year) course portfolio of the School of Information Sciences at the University of Pittsburgh.

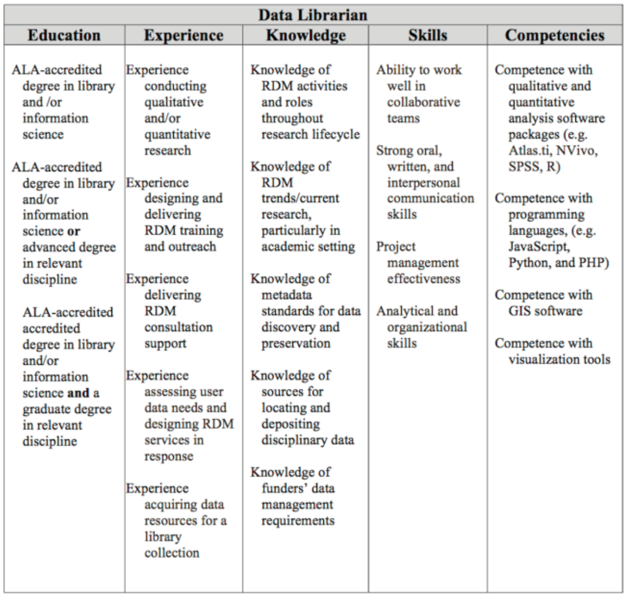
This analysis is followed by development of specific data-centric model pathways, composed of course "stepping stones." These pathways were developed based on the analysis of job posting requirements and through a review of the current (2014-2015 AY) iSchool course portfolio for Library and Information Sciences. We extended our review to include a consideration of courses in the Information Sciences program at the iSchool at Pitt and in other units on campus to consider stepping stones that may exist outside of the MLIS program as it stands.

The analysis of Key Requirements for each of the three roles (Data Librarian, Data Steward/Curator and Data Archivist) are presented in Tables 2-4. Of these Key Requirements, four were required by all the roles: a) Experience or knowledge or understanding of the researcher perspective, b) Knowledge of metadata standards and schema for data, c) Competence with statistical / analysis software packages and d) Knowledge of disciplinary data.

5.1 Data Librarian

The Data Librarian jobs invite candidates with, at minimum, a graduate degree from an ALA-accredited library and information science program (or an equivalent degree). For tenure-stream faculty librarian positions examined, there was a desire for applicants with a second graduate degree. Notable in the narrative in the job adverts is an interest in candidates who understand the researcher perspective from their experiences as researchers and who are committed to user-centered library services and resources.

Table 2. Key Requirements for Data Librarian



Data Librarian				
Education	**Experience**	**Knowledge**	**Skills**	**Competencies**
ALA-accredited degree in library and /or information science ALA-accredited degree in library and/or information science **or** advanced degree in relevant discipline ALA-accredited accredited degree in library and/or information science **and** a graduate degree in relevant discipline	Experience conducting qualitative and/or quantitative research Experience designing and delivering RDM training and outreach Experience delivering RDM consultation support Experience assessing user data needs and designing RDM services in response Experience acquiring data resources for a library collection	Knowledge of RDM activities and roles throughout research lifecycle Knowledge of RDM trends/current research, particularly in academic setting Knowledge of metadata standards for data discovery and preservation Knowledge of sources for locating and depositing disciplinary data Knowledge of funders' data management requirements	Ability to work well in collaborative teams Strong oral, written, and interpersonal communication skills Project management effectiveness Analytical and organizational skills	Competence with qualitative and quantitative analysis software packages (e.g. Atlas.ti, NVivo, SPSS, R) Competence with programming languages, (e.g. JavaScript, Python, and PHP) Competence with GIS software Competence with visualization tools

The Data Librarian positions have a unique focus on a) knowledge of research funding agency data management requirements, b) knowledge of RDM activities and roles throughout the research lifecycle and c) experience of designing and delivering research data management (RDM) training and outreach (Table 2). Navigating data-centric model pathways through the current MLIS course portfolio, we propose the following course "stepping stones" for prospective Data Librarians. Together, these stepping stones form curricular pathways.

Essential course stepping stones for a prospective Data Librarian will include:

- *Research Data Management*
- *Research Data Infrastructures*
- *Metadata*
- *Academic Libraries*
- *Preserving Digital Collections*
- *Research methods in LIS.*

Desirable course stepping stones will include:

- *Intro to Information Technologies*
- *Managing & Leading Information Services*
- *Digital Repositories (new course already in development)*
- *GIS for Librarians*
- *Information Visualization.*

Course development and collaborative partnership opportunities have been identified:

- *Programming for Librarians (new course already in development)*
- *Intro to Statistical Data Analysis / Data Analytics* (from Graduate Program in Information Science & Technology IST colleagues within the School).

5.1 Data Steward/Data Curator

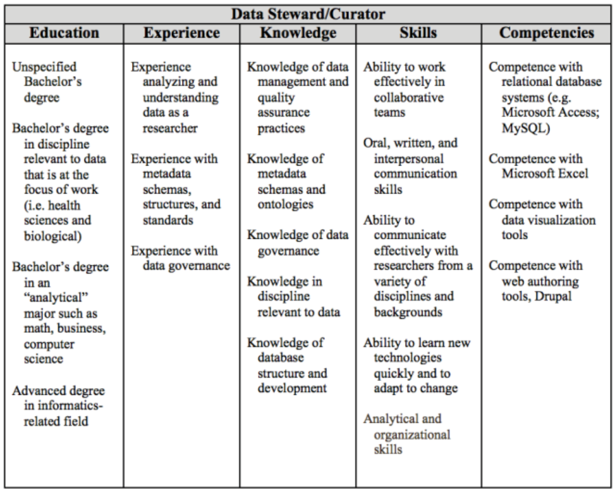
The Data Steward/Data Curator positions invited applications from individuals with a much wider range of disciplinary training. In addition to information science or library and information science degrees with course work in data modeling and metadata, employers were interested in candidates with computer science, mathematics, and business-related qualifications.

We identified a trend in job titles through our data collection. A search of indeed.com on March 30, 2015, produced 262 results that included the string "data steward." Conversely, there were only five results for "data curator," suggesting that the former, in the United States, is the more common position descriptor.

The Data Steward/Data Curator positions have a unique focus on a) experience of data governance and b) knowledge of data quality assurance practices and c) competency with relational database systems (Table 3).

Our search of the job banks involved using "data steward" and "data curator" as our search terms. What returned to us were positions that were both in the corporate and academic sectors. In the case of the former, these are positions that are data-centric but where the data is more likely to be used for internal research and compliance within the organization.

Table 3. Key Requirements for Data Steward/Curator Positions



Data Steward/Curator				
Education	**Experience**	**Knowledge**	**Skills**	**Competencies**
Unspecified Bachelor's degree Bachelor's degree in discipline relevant to data that is at the focus of work (i.e. health sciences and biological) Bachelor's degree in an "analytical" major such as math, business, computer science Advanced degree in informatics-related field	Experience analyzing and understanding data as a researcher Experience with metadata schemas, structures, and standards Experience with data governance	Knowledge of data management and quality assurance practices Knowledge of metadata schemas and ontologies Knowledge of data governance Knowledge in discipline relevant to data Knowledge of database structure and development	Ability to work effectively in collaborative teams Oral, written, and interpersonal communication skills Ability to communicate effectively with researchers from a variety of disciplines and backgrounds Ability to learn new technologies quickly and to adapt to change Analytical and organizational skills	Competence with relational database systems (e.g. Microsoft Access; MySQL) Competence with Microsoft Excel Competence with data visualization tools Competence with web authoring tools, Drupal

Essential course stepping stones for a prospective Data Steward/Curator will include:

- *Metadata*
- *Research Data Management*
- *Research Data Infrastructures*
- *Information Storage & Retrieval*
- *Digital Repositories (new course already in development)*
- *Preserving Digital Collections*

- *Information Architecture*
- *Corporate knowledge practices*
- *Database Management.*

Desirable course stepping stones will include:

- *Information Security & Privacy*
- *Data Structures*
- *Advanced Topics in Database Management*
- *Information Visualization*
- *Foundations of clinical & public health informatics (if interested in stewardship positions in health)*
- *Digital Curation.*

Course development and collaborative partnership opportunities have been identified:

- *Programming for Librarians (new course already in development)*
- *Data Governance (with Business School or School of Law).*

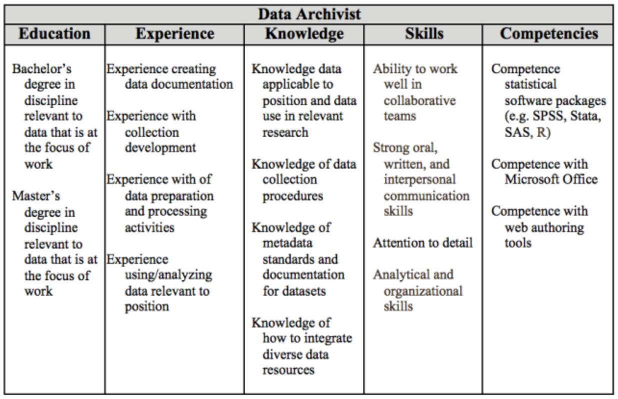
5.2 Data Archivist

We found a scarcity in current "data archivist" positions; as a result we were only able to code a small set of employer requirements for positions titled "data archivist" and with data archival responsibilities. This is probed further in our Discussion.

In addition, analysis revealed that the term "digital archivist" was out of scope to our analysis. The current job positions with this title that we located were records-focused and did not include any explicit mention of data as an information object under the purview of the candidate. There is, of course, an argument to be made that data is meaningful documentation for research and that, as such, all archivists are data archivists. For the purposes of this paper, we were primarily focused on structured data and as such did not cast our net to include positions without explicit allusion to this.

The Data Archivist positions have a unique focus on a) Experience of data documentation, b) Experience of data preparation and c) Knowledge of how to integrate diverse data resources (Table 4).

Table 4. Key Requirements for Data Archivist Positions



Data Archivist				
Education	**Experience**	**Knowledge**	**Skills**	**Competencies**
Bachelor's degree in discipline relevant to data that is at the focus of work Master's degree in discipline relevant to data that is at the focus of work	Experience creating data documentation Experience with collection development Experience with of data preparation and processing activities Experience using/analyzing data relevant to position	Knowledge data applicable to position and data use in relevant research Knowledge of data collection procedures Knowledge of metadata standards and documentation for datasets Knowledge of how to integrate diverse data resources	Ability to work well in collaborative teams Strong oral, written, and interpersonal communication skills Attention to detail Analytical and organizational skills	Competence statistical software packages (e.g. SPSS, Stata, SAS, R) Competence with Microsoft Office Competence with web authoring tools

Essential course stepping stones for a prospective Data Archivist will include:

- *Research Data Management*
- *Research Data Infrastructures*
- *Metadata*
- *Archives & Records Management*
- *Archival Appraisal*
- *Library & Archival Computing*
- *Preserving Information*
- *Preserving Digital Collections*

Desirable course stepping stones will include:

- *Access Systems, Standards, and Tools*
- *Digital Repositories (new course already in development)*
- *Preserving Digital Culture (course looking at historical development of digital media and theory of digital preservation)*
- *Digital Curation*
- *Information Architecture*
- *Web archiving*

Course development opportunities may encompass:

- *Intro to Statistical Data Analysis / Data Analytics* (from Graduate Program in Information Science & Technology IST colleagues within the School).
- *User experience and systems evaluation.*

6. DISCUSSION

This methodology utilizes the textual analysis of job postings for the three specific data roles, and has proved to be effective in revealing the particular qualifications, experience, knowledge, skills and competency requirements which employers are seeking from graduate students. By using only recent job descriptions, we are seeing the current perspective from the marketplace, where a wide range of organizations are recruiting to fill new data-centric positions. The associated mappings to graduate curriculum components gives an indication of the scope and contribution of the current course portfolio, and its relevance to a translational / market-driven data science environment.

6.1 Comparing the data roles

Previous analyses of data curation / digital curation job descriptions have highlighted a range of job titles with "archivist" featuring as a frequent term [32], great variation in job duties with corporate or research organizations more likely to require domain expertise, most often in science [33]. Job analyses have been used to determine health science and science and technology librarians' competencies for data management [34] and a set of digital curation competencies [35].

The sample size of jobs analyzed in this study was relatively small compared to the previous studies, but was targeted at very specific roles taking a "snapshot" approach. Our results suggested some common requirements of employers across the three job types. Employers were seeking graduates with experience or understanding of the researcher perspective; this knowledge could be attained by carrying out research as a doctoral student or by working closely with a research team, and indirectly implies demonstrating domain expertise. This requirement also emphasizes the value of immersive or embedded sessions or practicums in a research laboratory or other research environment within the curriculum, where the research lifecycle can be observed first-hand through bench science, experimental workflows, field work and the day-to-day perspectives and motivations of researchers exposed.

The common requirement for knowledge of disciplinary data may be related to acquiring research experience in a particular domain. Disciplinary perspectives can also be attained through immersive classes, skills labs, intensive hands on practicums, or internships [31]. Knowledge of metadata standards was an additional common requirement of all three roles and reflects the need for structured data descriptions using established domain schema wherever possible. An understanding of metadata issues also addresses the need for curators to understand the effort/cost : benefit balance for producers (creators) and consumers (re-users) of data [36].

The final common requirement was competence with statistical and/or analysis software packages such as SPSS, R, Excel and Stata. This is not particularly unique to data science roles or an absent requirement at many iSchools; there are indeed many iSchool programs that already require quantitative skills, including statistics and programming. The results from this study rather demonstrate the continuing relevance of quantitative skills in iSchool graduate students.

Each of the three data roles sought unique requirements for applicants. The Data Librarian roles required ALA-accredited qualifications, with selected requirements for a higher degree; there was clear weight given to evidence of educational trajectories towards careers in library and information work. The Data Librarian roles were unique in stating a need for knowledge of research funding agency data management requirements. This links to the documented development of Research Data Services in academic libraries which focus on providing advocacy on funder policy e.g. US National Science Foundation, National Institutes of Health policy statements for data management or data sharing plans as components of research proposals. The development of data management planning (DMP) tools such as the DMPTool[2] and DMPOnline[3] has enabled libraries to provide consultation and training services as elements within designated "data librarian" roles.

A further unique requirement was a stated focus on RDM activities throughout the research lifecycle. Whilst there are many representations of the research data lifecycle [37], [38], at each stage there are interventions where a data librarian can make a positive contribution e.g. recommending a metadata schema for data description, promoting an established and trusted repository for data deposit and assisting with data identification and citation processes. Therefore a thorough understanding of the whole data lifecycle opens up opportunities for data librarians to craft new data services to support the research community.

The final unique component for data librarians was an emphasis on RDM training and outreach. Academic libraries have long established working relationships with faculty and (graduate) students in departments and schools; frequently this relationship is enacted through liaison / faculty / subject librarians who develop and deliver outreach and training on aspects of information literacy, e-journals, open access publications etc. There is now a significant need for scaling up advocacy and outreach for the many components of RDM; some academic libraries are developing new Research Services portfolios which bring together a mix of novel RDM and digital scholarship offerings delivered by data librarians and others, working in newly sculpted research support teams. This trend may be accompanied by organizational restructuring to optimize resources, service functions and communications.

Reviewing Data Steward / Curator positions, we found some commonalities with the other roles, but also some unique features. These roles sought applicants from a wider range of disciplines and backgrounds. The reference to computer science, mathematics and business-related qualifications positioned these roles more closely with Data Engineer and Data Analyst roles. The trend towards using the term "data steward" perhaps reflects the strong profile of the National Agenda for Digital Stewardship promulgated by the National Digital Stewardship Alliance [39].

The Data Steward / Curator roles also demonstrated unique components: an emphasis on data governance which recognizes the importance of intellectual property rights and other legislative issues associated with data sharing and compliance with open data policy aspirations at both national and institutional levels. These roles also required a knowledge of data quality assurance practices, which reflect the key role of data curators in data selection, appraisal and cleansing workflows as critical elements of data ingest into (trusted) repositories and data centers. The third unique requirement was competency with relational database systems; many large-scale datasets are stored in very large and complex database systems with hundreds of columns and many rows. The ability to import, manipulate, export and manage data in such complex systems is essential in the era of "big data".

Our search for recent positions titled "Data Archivist" was challenging, with a significant lack observed in the particular job banks we trawled. We can speculate that whilst the archives community is currently recruiting digital archivists, in most cases, these roles draw on long-established traditions and terminology, and are not yet explicitly framed around data. This does not mean that these positions are not data-focused (primary resources are arguably data), but the absence of reference to datasets placed it outside of our methodology. A major finding of this study is that "Data Archivist" is an uncommon job title today. This points to the fact that archivist roles are being renamed and reclassified in different job sectors.

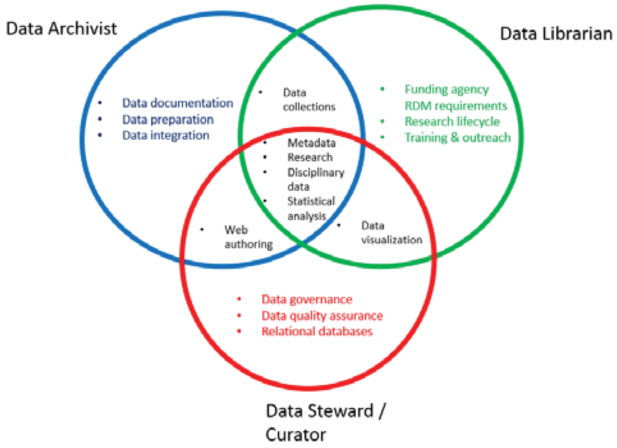
Considering the unique elements of the Data Archivist positions that were located, we identified an emphasis on "data documentation", "data preparation" and "data integration", which add weight to the assertion that established archival principles are reflected in the language which is applied to new digital objects of record i.e. research datasets. It can be argued that there are some similarities in the requirements for Data Librarian and Data Archivist; the emphasis on "data collections" is a common theme which once again reflects the long-established foundations of these fields.

Looking across the three roles, in general the requirements support the categorization and role descriptions of Lyon & Brenner [18], with some equivalence with the Data Librarian and Data Manager roles (the latter possibly equivalent to Data Steward/Curator) described by Pryor and Donnelly [19]. There are particular common Key Requirements across pairs of roles. The Data Archivist and Data Librarian positions emphasized experience related to collections (possibly demonstrating common foundational principles); the Data Archivist and Data Steward / Curator roles featured Web authoring competencies and the Data Librarian and Data Steward/Curator roles required competency with (Data) Visualization tools. The commonalities and unique aspects we observed are summarized in a Data Roles and Requirements Venn Diagram in Figure 1.

Figure 1. Data Roles and Requirements Venn Diagram

[2] DMPTool https://dmp.cdlib.org/

[3] DMPOnline https://dmponline.dcc.ac.uk/



6.2 iSchool Curriculum development

In reviewing our current curriculum, we are adopting a "model pathway" approach to reflect the optimal mix of courses (stepping stones) that a graduate student should/could take to follow specific career trajectories. This navigational process is primarily intended to guide the prospective student, but also serves to highlight potential opportunities to strengthen, broaden and extend the curriculum to support the breadth of data science roles described by Lyon and Brenner [17]. Whilst in some cases a truly radical re-engineering of the curriculum may be appropriate, rather we are adopting the approach of navigating the curriculum in new ways to signpost and showcase primary stepping stones (i.e. courses) to these emerging data science roles.

However given the specific requirements for the three roles explored in this small-scale study, we do suggest that the absence of any data-centric courses would be a perceived gap in an iSchool Library and Information Science curriculum at this time. Furthermore, it is clear that selected new elements are needed to fully meet the expectations of employers seeking data talent. These components may be acquired from other internal iSchool programs such as Information Science & Technologies at Pittsburgh, or from particular external sources e.g. other Schools and Departments. The current focus on "data" opens up many opportunities for new and exciting collaborations and partnerships in curriculum development.

The observed emphasis on disciplinary knowledge and experience may also be addressed through partnerships. For example, a new joint appointment with the Department of Biomedical Informatics paves the way for new focused offerings on text mining and data extraction, ontology development and knowledge organization systems (KOS). The diversity in disciplinary data practice is exemplified by the plethora of standards, schema, formats and cultures in different domains. As educators developing the data curriculum, ensuring graduate student expertise in all these fields is challenging; some would say impossible. Our approach is to aim for balancing these poles (knowledge of all data domains versus knowledge of none), within the curriculum through a mix of RDM courses which heavily feature case studies and domain exemplars, immersive sessions in the research laboratory and discipline/data-type-specific courses e.g. GIS for Librarians, Health Informatics. The results of this study validate the embedded / immersive / practicum components of data courses, since employers have stated their desire for applicants to demonstrate an understanding of research practice and disciplinary expertise in job postings.

The study also highlights the need to upskill existing practitioners as well as to produce data-savvy and work-ready graduates. Cohorts in the RDM and RDI courses have included practising librarians from both the Pittsburgh University Library System and Health Libraries System, and from Carnegie Mellon University Libraries. In this way, capacity and data capability is being scaled-up to meet the growing demand for Research Services. We hope to see a similar trend with graduates seeking careers within archival science and practicing archivists joining RDM/RDI courses.

This study raises some more general implications for recruitment strategies to graduate Library and information science courses. The employer demand for research experience, disciplinary knowledge and data analysis competencies, highlights the need for LIS programs to review their recruitment base to include STEM graduates who have strong technical and quantitative skills, and are happy manipulating tabular data or performing statistical analyses of datasets using a software package such as R or SPSS.

7. Conclusions

Our study has demonstrated key commonalities and distinct differences between the three data roles investigated. We acknowledge that the work is relatively small in scale and has a strong US focus, but the results indicate helpful directions for developing the iSchool curriculum to help to fill the data "talent gap" [13]. The translational data science approach adopted in the methodology (from iSchool to marketplace), reflects the trends of employers across sectors who are seeking data-savvy and work-ready graduates to fill these different data roles. Finally, we believe there is a great opportunity for iSchools to develop and extend their curriculum to embrace additional data-centric programs, courses and certificates to both educate new-entrants and to upskill existing practitioners to achieve the data-savvy profile, which is currently in high demand.

8. REFERENCES

[1] Woolf, S. H. (2008) The Meaning of Translational Research and Why It Matters. JAMA 299 (2), 211-213, Accessed April 13, 2015 http://jama.jamanetwork.com/article.aspx?articleid=1149350

[2] Carlson, S. 2006. Lost in a Sea of Science Data. *The Chronicle of Higher Education* (June 23). https://chronicle.com/article/Lost-in-a-Sea-of-Science-Data/9136

[3] Hey, T. and Hey. J. 2006. E-Science and its implications for the library community. *Library Hi Tech*, 24(4), 515-528. http://www.emeraldinsight.com/doi/pdfplus/10.1108/07378830610715383

[4] Macdonald, S. and Martinez-Uribe, L. 2008. Data librarianship – a gap in the market. *CILIP Update*, 7(6), 20-21. https://www.era.lib.ed.ac.uk/bitstream/handle/1842/2499/Gap%20in%20the%20market.pdf?sequence=3&isAllowed=y

[5] Corrall, S. 2012. Roles and responsibilities: libraries, librarians and data. In Pryor, G. (Ed.) Managing Research Data. Facet Publishing. pp105-133.

[6] Lyon, L. 2012. The Informatics Transform: Re-engineering Libraries for the Data Decade. *IJDC* 7(1), 126-138. http://www.ijdc.net/index.php/ijdc/article/view/210/279

[7] Jaguszewski, J. M. and Williams. K. 2013. New roles for new times: transforming liaison roles in research libraries. ARL Report. http://www.arl.org/storage/documents/publications/nrnt-liaison-roles-revised.pdf

[8] Auckland, M. 2012. Re-skilling for Research. Research Libraries UK (RLUK) Report. http://www.rluk.ac.uk/wp-content/uploads/2014/02/RLUK-Re-skilling.pdf

[9] Cox, A. 2012. Upskilling Liaison Librarians for research Data Management. *Ariadne* (6 December) http://www.ariadne.ac.uk/print/issue70/cox-et-al

[10] Tenopir, C., Birch B. and Allard, S. 2012. Academic libraries and Research Data Services: Current Practices and Plans for the Future. ACRL White Paper. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf

[11] Cox, A. and Pinfield. S. 2014. JoLIS 46(4), 299-316. http://lis.sagepub.com/content/46/4/299.full.pdf+html

[12] HM Government 2013. Seizing the data opportunity. A strategy for UK data capability. White Paper. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/254136/bis-13-1250-strategy-for-uk-data-capability-v4.pdf

[13] Bakhshi, H., Mateos-Garcia J and Whitby, A. 2014. Model workers: How leading companies are recruiting and managing their data talent. Nesta Report. http://www.nesta.org.uk/sites/default/files/model_workers_web_2.pdf

[14] Lyon, L. 2007. Dealing with Data: Roles, Rights, Responsibilities and Relationships. Consultancy Report. http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf

[15] Swan, A. and Brown, S. 2008. The skills, role and career structure of data scientists and curators: an assessment of current practice and future needs. Report to the JISC. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.8960&rep=rep1&type=pdf

[16] Larsen, R. 2014. What can we learn from the data? Preparing the workforce for digital curation: The iSchool perspective. International Digital Curation Conference. Presentation available from http://www.dcc.ac.uk/sites/default/files/documents/IDCC14/Panels/RonLarsen_Panel.pdf

[17] Varvel, V. E. et al 2010. Report from the Research Data Workforce Summit. https://www.ideals.illinois.edu/bitstream/handle/2142/25830/RDWS_Report_Final.pdf

[18] Lyon, L. and Brenner, A. 2015. Bridging the Data Talent Gap – positioning the iSchool as an Agent for Change. *IJDC* 10(1), 111-122. http://www.ijdc.net/index.php/ijdc/article/view/10.1.111/384

[19] Pryor, G. and Donnelly, M. 2009. Skilling up to do data: Whose role, whose responsibility, whose career? *IJDC* 4(2), 158-170. http://www.ijdc.net/index.php/ijdc/article/view/126/133

[20] Creamer, A.T., Morales, M.E., and Kafel, D. 2012. A sample of Research Data Curation and Management Courses. *J. eScience Librarianship* 1(2), 88-96. http://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1016&context=jeslib

[21] Cox, A., Verban, E. & Sen, B. (2014) A Spider, an Octopus or an animal just coming into existence? Designing a curriculum for librarians to support research data management. *J eScience Librarianship* 3(1) 15-30, accessed April 15, 2015. http://escholarship.umassmed.edu/jeslib/vol3/iss1/2/

[22] Shadbolt, A., Konstantelos, L., Lyon, L. and Guy, M. 2014. *IJDC* 9(1), 313-323. http://www.ijdc.net/index.php/ijdc/article/view/9.1.313/360

[23] Macdonald, S. and Rice, R. 2012. "DIY" research Data management Training Kit for Librarians. Available from http://ceur-ws.org/Vol-1016/paper27.pdf

[24] Tibbo, H. R. 2015. Digital curation education and training: From digitzation to graduate curricula to MOOCs. *IJDC* 10(1) 144-153. http://www.ijdc.net/index.php/ijdc/article/view/10.1.144/387

[25] Carlson, J. R. et al 2011. Determining Data information literacy needs: a study of students and research faculty. *Portal: Libraries and the Academy* 11(2), 629-657. http://muse.jhu.edu/journals/portal_libraries_and_the_academy/v011/11.2.carlson.pdf

[26] Carlson, J. et al, 2013. Developing an Approach for data management education: a report from the Data Information Literacy Project. *IJDC* 8(1) 204-217. http://www.ijdc.net/index.php/ijdc/article/view/8.1.204/306

[27] Molloy, L. and Snow, K. 2011. DaMSSI Project Final Report. Available from http://www.academia.edu/2808837/DaMSSI_Data_Management_Skills_Support_Initiative_Final_Report

[28] Rubio, D. M. et al 2010. Defining Translational research: Implications for Training. *Acad Med*. 85(3) 470-475. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2829707/

[29] Lyon, L. and Webster, K. 2014. Embedding immersive informatics research data management within the iSchool curriculum: a laboratory-based action research Case Study. *Library Research Seminar VI,* University of Illinois Urbana-Champaign. Abstract available at http://www.library.illinois.edu/lrs6/Library_Research_Seminar_VI_Program.pdf

[30] Martin, E. R. (2013) Highlighting the Informationist as a data librarian embedded in a research team. (Editorial) *J eScience Librarianship* 2(1) 1-2, accessed April 13, 2015. http://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1044&context=jeslib

[31] Mayernik, M. S. et al 2015. Enriching education with exemplars in Practice: Iterative development of data curation internships. *IJDC* 10(1) 123-134. http://www.ijdc.net/index.php/ijdc/article/view/10.1.123/385

[32] Lee, C. (2008) What do Job Postings indicate about Digital Curation Competencies? Presentation at the Society of American Archivists Research Forum, August 26, 2008, San Francisco, accessed April 13, 2015. http://ils.unc.edu/digccurr/digccurr-saa-research-forum-2008.pdf

[33] Cragin, M. H. et al (2009) Analyzing Data Curation Job Descriptions. Poster: 5th Int. Dig. Curation Conference, London, 2-4 December 2009, accessed April 13, 2015. https://www.ideals.illinois.edu/bitstream/handle/2142/14544/Cragin_poster_abstract_DCC_09.pdf?sequence=2

[34] Creamer, A., Morales, M & Crespo, J. et al (2012) An assessment of needed competencies to promote the data curation and management librarianship of health sciences and science and technology librarians in New England. *J of eScience Librarianship,* 1(1) 18-26. http://escholarship.umassmed.edu/cgi/viewcontent.cgi?article=1006&context=jeslib

[35] Kim, J., Warga, E. & Moen, W. (2013) Competencies required for digital curation: An analysis of Job Advertisements. IJDC 8(1) 66-83, accessed April 13, 2015. http://www.ijdc.net/index.php/ijdc/article/view/8.1.66

[36] Michener, W.K. et al (1997) Non-geospatial metadata for the Ecological Sciences. *Ecological Applications* 7(1), 330-342, accessed April 13, 2015. http://lits.bio.ic.ac.uk:8080/litsproject/Michenereta11997.pdf

[37] Higgins, S. (2012) The Lifecycle of data management. Chapter 2 in Ed. Graham Pryor. Managing Research Data Facet Publishing, 239pp.

[38] Corti, L, et al (2014) The Research Data Lifecycle. Chapter 2 in Managing and Sharing Research Data. A Guide to Good Practice. Sage Publications, 222pp.

[39] NDSA (2014) National Agenda for Digital Stewardship 2015 accessed April 13, 2015. http://www.digitalpreservation.gov/ndsa/documents/2015NationalAgenda.pdf

A Survey of Organizational Assessment Frameworks in Digital Preservation

Emily Maemura, Nathan Moles, Christoph Becker
Faculty of Information
University of Toronto
{e.maemura, n.moles}@mail.utoronto.ca, christoph.becker@utoronto.ca

ABSTRACT

As the field of digital preservation continues to mature, there is an increasing need to systematically assess an organization's abilities to achieve its digital preservation goals. A wide variety of assessment tools exist for this purpose. These range from light-weight checklists to resource-intensive certification processes. Conducted as part of the BenchmarkDP project, this paper presents a survey of these tools that elucidates available options for practitioners and opportunities for further research.

General Terms

Institutional opportunities and challenges; Frameworks for digital preservation

Keywords

capability, maturity, risk, organizational assessment, design science

1. INTRODUCTION

Over the past two decades digital preservation (DP) research has produced a wide range of tools, models, strategies, and other innovations to facilitate the long-term management of digital objects. Although much progress has been made in this area, solutions targeting individual components do not work in isolation and consideration must be given to digital preservation capabilities at the organizational level. Unfortunately, the DP community currently lacks standardized assessment tools to facilitate rigorous and systematic evaluation of an organization's capacity to achieve its preservation goals. Systematic assessment at the organizational level is essential to evaluate the efficacy of an organization's DP operations, to provide reliable benchmarks against which continuous improvement can be made, and to enable comparisons across institutions.

The BenchmarkDP project is developing and evaluating rigorous, systematic, and evidence-based means for comparing techniques, approaches, tools, and systems. As part of the project's ongoing study of organizational assessment in DP, this paper provides a comprehensive survey of existing models and frameworks that assess an organization's ability to achieve its DP goals through a combination of people, technology, and processes. The survey is driven by two research questions: *(RQ1) What are the options for organizational assessment, and how do they vary in terms of focus, requirements, and expected outputs? (RQ2) What trends and gaps exist in the current landscape, and do these present opportunities for research?*

In answering these questions, the survey will aid practitioners in comparing the different options for organizational assessment, including the strengths and limitations of each approach. As well, this work will outline potential new directions for researchers and highlight areas where further study is needed. The first sections of this paper provide a brief background in assessment and key concepts, the rationale for the selection of the models surveyed, and a brief description of each model. A more detailed analysis and discussion follows. A concluding section highlights gaps in the current spectrum of solutions and identifies opportunities for further research.

2. BACKGROUND

The long-term focus of DP requires a set of coordinated activities and supporting infrastructure that includes people, technology, systems, information, and processes. This work is carried out by an organization (or an organizational unit that is part of a larger body) with the responsibility of preserving and providing access to digital information. As the field of DP matures, more systematic methods of understanding and comparing these activities are needed in order to assess the current state of preservation capabilities, identify areas that need improvement, and direct improvement efforts. Organizational assessment provides a method of measuring current performance and enables steps towards increased capacity, improved reliability, demonstrated trustworthiness, or reduced risk.

Outside of DP, this challenge of organizational assessment has been approached in different ways. We focus here on maturity models, as they are a prominent means of systematic assessment in other fields, with existing foundations to draw on. Maturity models generally, and Capability Maturity Models specifically, can be used to take an informed approach to continuous improvement[29].

These concepts originate in the Software Engineering Institute Capability Maturity Model (CMM) developed to ensure reliable and consistent processes within the field of Software Engineering [35]. While many models and frameworks have been developed based on this original CMM, there is still little consensus or consistency in the meanings and uses of the terms 'capability' and 'maturity' [25]. We define capability broadly as the sustained ability to achieve a goal, through a combination of people, technology, and process[7]. Maturity is more difficult to define. Others have noted the different senses or aspects of maturity that are often confused [25], but all stem from the common dictionary definition of 'a state of completeness.' What is of primary interest for assessment is the process of bringing something to maturity, the path to completeness [25]. To achieve this, maturity models describe the different sequential stages of growth – an 'evolutionary path' – that target individual processes or multiple dimensions. An organization's overall state of maturity provides a measure of how much confidence one can have in the organization to successfully achieve goals and consistently provide services – in short, the degree of reliability and predictability.

Recent work demonstrates a growing interest in assessment through maturity model frameworks in other domains[50], and we see this growing interest mirrored in DP. While we include many models in our survey that are not formal CMMs and do not use the vocabulary of maturity models generally, we maintain that both of these dimensions (the capabilities available, and the predictability in successfully achieving goals) must be addressed for organizations to meet the challenges of DP. We will therefore use these concepts, and the associated literature on maturity models from other domains, to examine frameworks for organizational assessment.

In describing these existing approaches to organizational assessment in the domain of DP, we will discuss models, methods, tools, and frameworks. For our purposes, **models** are a "formal description of some aspects of the physical or social reality for the purpose of understanding and communicating"' (Mylopoulos, quoted in [29]). Mettler describes **methods** as "systematic (i.e. they deliver rules on how to act and instructions on how to solve problems), goal-oriented (i.e. they stipulate standards on how to proceed or act to achieve a defined goal), and repeatable (i.e. they are intersubjectively practicable)" [29]. A **tool** is a concrete or abstract thing used to perform a task[1]. Finally, a **framework** is the overall set of components, including at a minimum a model, and any associated methods or tools.

3. OVERVIEW OF FRAMEWORKS

Many assessment frameworks are referenced in discussions of DP. We cast a wide net for this survey, with literature searches in Scopus and Google Scholar for permutations of 'digital preservation' and 'tool,' 'framework,' 'model,' 'capability,' 'maturity,' 'measurement,' 'improvement,' and 'assessment,' as well as snowball sampling of the citations from this initial set of literature. Community venues and websites were explored, such as the Preservation and Archiving Special Interest Group (PASIG), the Digital Preservation Coalition, and the blog *The Signal*[2]. Additional models were included based on our own familiarity with existing publications.

Table 1: Introducing the different Frameworks

Name	Abbreviation	Year
The Five Organizational Stages of Digital Preservation [26]	Stages	2003
Capability Assessment and Planning Toolkit [39]	CTG	2005
DRAMBORA [27]	DRAMBORA	2007
JISC AIDA Toolkit [37]	AIDA	2008
Data Seal of Approval	DSA	2010
ISO16363 / TRAC [4]	ISO16363	2012
Digital Preservation Capability Maturity Model [19]	DPCMM	2012
SHAMAN capability model [6]	SHAMAN	2012
Maturity Levels [13]	Brown	2013
NDSA Levels of Digital Preservation	Levels	2013
DIN31644 / NESTOR Seal [2]	NESTOR	2013
NSLA Maturity Matrix [36]	NSLA	2013
Scoremodel	Scoremodel	2013
e-ARK [41, 40]	e-ARK	2015

We then developed a set of inclusion and exclusion criteria. First, the assessment must be explicitly focused on the domain of DP. Many models address closely related domains such as Records Management or Information Governance. However, maturity models by JISC and ARMA as well as the ECMMM[3] were excluded since they do not address concepts or concerns specific to DP such as those outlined in OAIS[3] or TDR[42]. Similarly, the CMM for Scientific Data Management [16] was excluded as it addresses concerns specific to research data, and only covers DP from a high-level perspective.

Further, within the domain of DP, we included only models that target the organization (or organizational unit). We excluded the SPOT model for Risk Assessment[49], Data Curation Profiles Toolkit[4], and the Data Asset Framework[5] as they target a specific function only. The PLATTER framework[18] and NEDCC checklist[12] were also excluded as they cover initial planning but not systematic assessment for improvement. The Preservica DPMM[6] was excluded since it targets storage media, or storage services. Finally, practical criteria were considered — the model and assessment framework must be freely available online, and in English. Due to language barriers, the Dutch ED3[7] and the German DIN31644 standard were excluded.

In total, 14 models met all criteria, described briefly below in chronological order and listed (with abbr.) in Table 1.

Five Organizational Stages of Digital Preservation (2003) – This is the earliest example of a model for organizational assessment and improvement focused on DP. Its motivation stems from an attempt to shift discussions away from technologically oriented solutions, towards 'or-

[1] Oxford English Dictionary "tool, n." www.oed.com

[2] http://blogs.loc.gov/digitalpreservation/

[3] https://ecmmaturity.files.wordpress.com/2009/02/ec3m-v01_0.pdf

[4] http://datacurationprofiles.org/

[5] http://data-audit.eu/

[6] http://www.preservica.com/download/852

[7] http://www.den.nl/standaard/225/

ganizational response and readiness' issues. The target audience is defined broadly as all 'cultural repositories,' but examples used draw heavily on work with research libraries (mostly the authors' experiences at Cornell University Library). A community-created model, the structure is simple and lightweight, with three key indicators (policy and planning; technological infrastructure; content and use) for each of the five stages. It uses a conception of DP based on the three-legged stool model: organizational infrastructure, technological infrastructure, and resources framework.

Center for Technology in Government (CTG) Capability Assessment and Planning Toolkit (2005) – This model was released by the CTG at SUNY Albany. Built on the basis of the UNESCO Guidelines for the Preservation of Digital Heritage[30] and the *Stages*[26], it was developed in collaboration with the Library of Congress, with input from the broader community. It is intended to guide self-assessments of the DP capabilities of state governments and government agencies, to be used by a range of librarians, archivists, records managers, and other information professionals. The assessment process, conducted through a series of workshops, aims to identify gaps and weaknesses in 19 dimensions of capability. The toolkit provides a range of useful templates and examples.

DRAMBORA (2007) – The Digital Repository Audit Method Based On Risk Assessment was created as a joint project of the Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE). This model approaches organizational readiness through risk assessment, complementary to other risk models that are not the focus of this survey [49, 10]. The 'internal audit' method progresses through 6 stages, beginning with documenting contextual information about the organization then identifying specific assets and activities, and risks, with probability and impact. A common framework of four Operational functions and four Support functions is used throughout the audit. There is an offline toolkit of templates (MS Word and Excel files), and an online form that streamlines the process and provides a summary report. Use of the online toolkit further allows for comparison with results from similar repositories that have completed the assessment.

JISC Assessing Institutional Digital Assets (AIDA) Toolkit (2008) – The AIDA Toolkit, created by the University of London Computer Centre, operationalizes the Stages through a self-assessment tool aimed at evaluating the digital asset management and DP readiness of higher education institutions in the UK [37]. The Toolkit Mark I was released in 2008 and a revised Mark II released in 2009. The objective of this toolkit is to capture an accurate picture of the current state of an organization's readiness and capabilities, not to provide explicit direction for improvement. The assessment process is based on a weighted score system that the AIDA project team requests in order to provide feedback. The toolkit contains templates and scorecards as well as an online tool.

Data Seal of Approval (DSA) (2010) – The DSA[8] is a simple list of criteria, and an online tool, created by Dutch-based DANS. It consists of 16 guidelines, in 3 categories: Related to Data Producers, Related to Repositories, and Related to Data Consumers. Each guideline receives a rank of 0-4 based on the five possible responses/statements, the highest being 'implemented.' The assessment is presented as a two-tiered model, supporting self-assessment for internal improvement and a 'seal of approval' for meeting the guidelines, verified through a peer-review process. Between the initial release in 2010, and the current version from 2013, 41 seals have been awarded. All material for conducting the assessment is freely available online, including manuals for applicants and reviewers. All of the documentation from the awarded seals is available on the DSA website.

ISO16363/TRAC (2012) – ISO16363:2012[4] is a standard for an audit process of the trustworthiness of digital repositories, based on compliance with ISO14721 (OAIS)[3]. It builds on the influential 2002 report which outlined the attributes and responsibilities of a Trusted Digital Repository[42] and the subsequent and the subsequent collaborative work between RLG, NARA, and CRL which resulted in the Trustworthy Repository Audit and Certification (TRAC) Criteria and Checklist[34] published in 2007. The audit process for the standard is lengthy and resource-intensive. It takes into account a wide range of organizational, infrastructure, security, and management factors. Certification is available through organizations such as the Centre for Research Libraries (CRL) and the Primary Trustworthy Digital Repository Authorization Body (PTAB), usually at considerable cost. Several repositories have been certified using this process, and there is some indication that the standard can be used for self-assessment. A tool[9] developed by MIT has been built on this standard. A standard outlining requirements for bodies providing certification based on ISO16363, has recently been released as ISO16919:2014.

DPCMM - Dollar & Ashley (2012) – This DP Capability Maturity Model was created by consultants Charles Dollar and Lori Ashley. The model was first made available in 2012, with an updated version released in 2014. Based loosely on other CMMs, the model defines five levels or stages of capability in digital preservation: Nominal, Minimal, Intermediate, Advanced, and Optimal. The requirements for each level are specified for 15 different components, covering both Infrastructure and Services. The model is largely based on OAIS, drawing heavily on the model's concepts and vocabulary from these standards and using compliance with ISO14721 as a threshold for certain stages. For each component a table is presented defining requirements to achieve each level, paired with a score from 0-4. Scores are then summed to provide the Aggregated Digital Preservation Capability Index Score. Additional reports, such as a 'road map,' are understood to be provided if the assessment is undertaken by the consultants. They have also introduced an online tool, available at `www.digitalok.org`.

SHAMAN capability model (2012) – The SHAMAN Reference Architecture[6], based on enterprise architecture concepts, includes a capability-based model of DP that identifies 11 capabilities in three groups: governance, risk, and compliance capabilities; business capabilities; and, support capabilities. The emphasis for preservation is on the four capabilities (acquire content, preserve bit streams, preserve

[8] `http://datasealofapproval.org/en/`

[9] `www.archivematica.org/wiki/Internal_audit_tool`

content, and disseminate content) that comprise the category of business capabilities and which are supported by the capabilities in the remaining categories. This capability model was further developed into a checklist assessment method that contains five steps: identify stakeholders, identify influencers, derive preservation goals, determine capabilities, assess capability level [8].

Adrian Brown's Maturity Levels (2013) – Presented in 2011 and subsequently published in a book [13], the development of this model was inspired by P2MM from the field of project management. No specific methodology is described for the assessment. Instead, each process perspective and capability level is paired with a statement about an action taken or process in place, such as "A written, approved digital preservation policy exists." No specific statements are provided for the lower levels of 'awareness,' so the model only provides a three-level scale of Basic Process, Managed Process, and Optimized Process. We have not encountered any applications of this model.

NDSA Levels of Digital Preservation (2013) – This model is a tiered matrix of practical recommendations created by the National Digital Stewardship Alliance.[10] Intended to assist organizations in establishing and improving DP activities, this model can also be used to assess the level of preservation achieved for specific digital objects or groups of digital objects. It was intended to achieve a middle ground between the complexity of ISO 16363 and simple guidance checklists [22]. The *Levels* consist of five dimensions and four progressive levels of maturity. With a focus on five functional areas at the operational level, the model is missing many of the organizational elements or dimensions present in other frameworks, but is nevertheless useful for certain applications.

DIN31644 / NESTOR Seal (2013) – Based on the NESTOR Catalogue of Criteria for Trusted Digital Repositories (2006)[32], the NESTOR Seal is an extended self-assessment process for digital archives, covering 34 criteria separated into three areas (organizational, technical, and infrastructure & security). It is based on the German DIN31644 'Criteria for trustworthy digital archives,' but structured in a way similar to the DSA, providing an extended reviewed self-assessment. An organization may apply for a NESTOR Seal that recognizes compliance with these criteria (but is not an accredited certification) by providing the documentation of their self-assessment for review by NESTOR. The review will be completed within three months; there is a 500 Euro fee for applying for the seal. It is envisioned as the middle-ground between the lightweight assessment of the DSA, and the intensive auditing of ISO16363. The full text of the DIN standard is available only in German, but the criteria have been summarized in English for the NESTOR seal in an 'Explanatory Notes' document [33].

NSLA Digital Preservation Environment Maturity Matrix (2013) – This model was created by the National & State Libraries of Australasia DP Group. Based on OAIS, this work aims to determine digital preservation maturity in relation to the OAIS Functional Entities through a five level CMM derived from the original SEI CMM[35]. Each Functional Entity is associated with one of the five levels during the assessment process through a series of questions completed by the auditor. The purpose is to identify the levels of maturity, development needs, and collaboration needs of the NSLA member institutions. No recommendations or guidelines for improvement are provided by the Maturity Matrix, however it includes assessment templates.

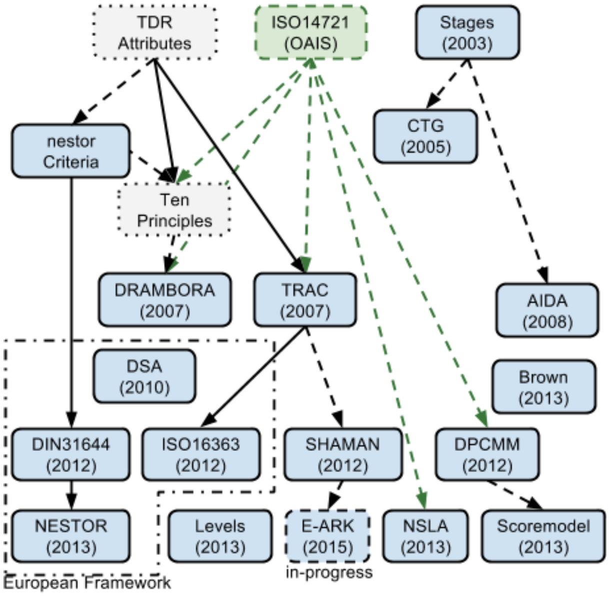


Figure 1: Family Tree of the different Frameworks.

Scoremodel (2013) – Scoremodel[11] is an online tool to identify risks and threats to digital objects, as well as provide basic recommendations. It is organized around seven clusters: organization and policy, preservation strategy, expertise and organization, storage management, ingest, planning and control, and access. In each of these sections, the tool presents users with a series of yes or no questions, each with context, associated risk and risk level, and an example of the evidence to be considered in the answer. Scoremodel is free to use online, open to all users, and available in both English and Dutch. However, the concepts, model, and rationale behind the tool are unclear and limited documentation is available.

E-ARK (2015) – In early 2015, the European Archival Records and Knowledge Preservation project released a maturity model for information governance that included many components outside the boundaries of this study. This work has continued with the release of an archiving maturity model, including an initial assessment and evaluation of a pilot study of 7 Archives released in October 2015. This model draws on TRAC and OAIS and presents a self-assessment questionnaire of 35 questions with responses corresponding to five levels of maturity. Questions are also grouped into five main capability areas: Pre-Ingest, Ingest, Archival Stoage and Preservation, Data Managment, and Access.

The relationships and influences of these models are mapped in a 'Family Tree' shown in Figure 1. This diagram also includes other influential documents and standards (indicated

[10] http://digitalpreservation.gov/ndsa/activities/levels.html

[11] http://scoremodel.org/en

by dashed boxes) that are not assessment models, specifically the OCLC/RLG attributes for Trusted Digital Repositories (TDR)[42], ISO14721/OAIS (2003, revised 2012), and the Ten Principles, developed jointly by CRL, DCC, DPE, and NESTOR (2007)[12]. Solid arrows show direct connections and evolution of models, while dashed arrows indicate explicit but loose influence.

4. ANALYSIS

4.1 Analytical Framework

An analytical framework is necessary to better understand this wide-ranging set of models, to find insightful patterns or trends. There is a growing body of literature studying maturity models in other domains, and we draw from this work in our analysis. In particular, the broad field of Information Systems has developed a rich body of knowledge on Design Science research methodology that approaches problem-solving through study and iteration of designed artifacts [23], and this previous work includes development of design principles in relation to maturity models and assessment frameworks.

Wendler[50] notes the variety of research that exists on maturity models, and we have attempted to cover a wide range to form our theoretical foundation here. First, the work by Jokela et al.[24] provide a similar survey of models in the domain of usability that focuses on the application of models. Second, to understand the models as artifacts, we have drawn on work in Design Science research, including examples and approaches define requirements for the process of developing a model [11], as well as general design principles for maturity models [38].

We determined a number of attributes to address our first research question regarding the existing options available for organizational assessment (and their focus, requirements, expected outputs). We first determined the **primary purpose**, understood here as the intended central aim of the model, and the motivation for undertaking the assessment. We defined three categories for primary purpose: initial planning, improvement, and certification. We also examined the nature of the assessment process and expected outputs. Specific requirements are necessary for different types of intended **audience** of the model, e.g. to be shared internally in the organization, with external stakeholders, or both. As well, we considered the **mode of application** (or 'method of application' in [17]), e.g. whether it is performed as a self-assessment, third-party assisted, or by a certified practitioner.

Next, we examined the degree of **concrete guidance** provided by each model, understood as the amount of clarity and documentation provided for applying the model, and the method of assessment [24]. This should also reflect that the method is 'systematic, goal-oriented, and repeatable' [29]. Here we extend this to include a discussion of the degree of detail or granularity provided in the results of the assessment and recommendations for improvement. We further noted which models provide formal documentation of **methods of assessment**, and what other **tools** are available for use. (Methods and tools are defined in Section 2 above). Additionally, **Empirical evidence** is used to describe if and how use of a model is substantiated [24]; we have used a broad definition here to consider evidence of any/all applications, including case studies.

Finally, we note that Wendler[50] draws a distinction between research conducted with maturity models and research conducted on or about maturity models. Research 'with' maturity models includes all research related to the development, application, and validation of a model. Research 'on or about' maturity models can be seen as the "meta" work that takes the maturity models themselves as the subject of research. One of the salient conclusions of Wendler's mapping study was that there was a need for further research 'on or about' models, and that the development of such work can have significant implications for both researchers and practitioners. Research on or about models would lead to fewer, but better (theoretically rigorous and empirically validated) models, discussed further in Section 5.

4.2 Analysis and Results

The analytical framework reveals a number of patterns and common traits of these models, and the overall results of the analysis are summarized in Table 2. Examining the intended **audience** and **purpose** of the models reveals that almost all 'planning' and 'improvement' models are intended primarily for internal audiences. Only three certification-oriented models (DSA, NESTOR, ISO16363) were found, and are all part of the European Framework for Audit and Certification of Digital Repositories[13] that reflects a path of progressively rigorous audits. While intended for external audiences, assessments with 'certification' models can also be used internally.

The majority of the models use self-assessment as the **mode of application**, though some pair self-assessment with third-party assistance, such as the peer-review methods of DSA, NESTOR, AIDA and NSLA. DPCMM is the only model using third-party assistance through a commercial consulting service. Brown and e-ARK provide no clear documentation of application, and are noted as N/A. ISO16363 is the only model intended for assessment by a certified practitioner, to be standardized through ISO16919 "Requirements for bodies providing audit and certification of candidate trustworthy digital repositories." [5]

A key finding of this analysis is that most models provide little **concrete guidance** for assessment or subsequent improvement measures. The 'certification' models provide more thorough documentation and the 'initial planning' models provide the least. A handful of models provide documents describing **methods of application** (DRAMBORA, SHAMAN, DSA, NESTOR). However, most provide little direction or instruction for application or result in recommendations for improvement (though commercial products like DPCMM may have a more detailed method of assessment that is not publicly available). Nine of the fourteen models provide some kind of **tool** to aid in carrying out the assessment. These range from paper-based templates to electronic forms to interactive online tools. Some online tools can be seen as providing an implicit step-wise method, however, this is not made explicit.

[12] http://tinyurl.com/qgnt367

[13] http://www.trusteddigitalrepository.eu/Memorandum%20of%20Understanding.html

Table 2: Surveyed models, methods and tools for organizational assessment in DP.

Name	Audience	Method	Tool	Mode	Concrete Guidance	Empirical Evidence
					Primary Purpose: Initial Planning	
Stages	internal	-	-	SA	Limited. Key indicators note high-level processes	Examples from Cornell, no further case studies.
Levels	internal	-	-	SA	Set of practical recommendations for use exists.	Content-based case studies
					Primary Purpose: Improvement	
CTG	both	-	PT	SA	Significant guidance for application through workshops, including template usage and data gathering.	Extent of use is unclear, limited evidence.
AIDA	internal	-	EF/ OT	SA (TPA)	Limited to instructions for tool; recommendations and feedback provided by project team	Multiple applications, but little documentation or evidence available
DRAM-BORA	internal	Y	EF/ OT	SA	Guidance documents are available, very detailed results	Extensive and well documented
Brown	internal	-	-	N/A	Limited. High-level processes identified	None
SHAMAN	external	Y	-	SA	No guidance on using the model	None
DPCMM	internal	-	OT	TPA	Limited to the description of the model	Model has been applied, but no documentation or evidence
NSLA	both	-	PT	SA (TPA)	Some guidance is provided for use of the tool. Results are limited to identifying areas of weakness	Only the initial study for which the tool was created
Score-model	internal	-	OT	SA	Limited recommendations both for use and in results	Some previous assessments can be seen.
e-ARK	internal	-	OT	SA	Limited to description of model	Results of pilot study available
					Primary Purpose: Certification	
DSA	both	Y	OT	SA, TPA	Guidance documents are available for applying for seal	Many applications, publicly available documentation, some published case studies
ISO16363	both	-	(OT)	all	Guidance documents are available for conducting audit	Many applications and case studies
NESTOR	both	Y	EF	SA, TPA	Little guidance beyond addressing documentation to provide for seal	Multiple applications, but limited evidence

Legend: PT= paper templates; EF=electronic forms; OT=online tools;
SA=Self-assessment; TPA=Third-party assisted; CP=Certified Practitioner; all = SA, TPA and CP

Further, many of the models are supported by no **empirical evidence** at all, with only very weak indicators of successful application (such as case studies), and no direct supporting evidence. The 'certification' models provide the greatest number of examples in terms of application documentation and case studies, but still provide little empirical evidence to establish user trust or demonstrate validity.

5. DISCUSSION

5.1 Trends and Tensions

This analysis of models and frameworks has generated a number of insights into both the larger field of DP and the models themselves, as well as shedding light on the tensions around systematic assessments modeled after CMMs.

One significant trend to emerge from this comparison is a marked increase in the number and complexity of models in recent years. This increasing interest in assessment models mirrors the increasing number of operational repositories and commercial offerings [46] and corresponds to the findings in a recent survey [14]. However, a greater number of models has not helped to address the challenges associated with assessment, and it is increasingly difficult to weigh the costs and benefits of different approaches. There are still tensions between standards-compliance and improvement, and balancing simplicity in carrying out the assessment with reliability or trustworthiness of the results. Reliability of results is often only achieved with significant investments of time, effort and cost. Even then, few models currently provide results that can be used to directly inform planning or decision-making; others note that the existence of such a decision mechanism for improvement paths is a fundamental design principle for prescriptive use of maturity models [38].

Many models use a numeric rating, and also translate each level into direct questions of which criteria are met. However, experience in process assessment has shown that translating criteria into questions does not result in accurate descriptive results. 'If you want to assess the maturity of a process, you do not take the direct approach of asking people whether they think the ... process is managed or established in their organization.'[9]. Similarly, challenges in finding consensus on ratings using direct questions are unsurprising. In fact, the SEI Appraisal Requirements for CMMI forbid the usage of numeric ratings if the assessment does not meet the stringent requirements of the highest-class assessment method[48].

Increased interest and development of assessment models can indicate the field's transition from a 'skilled artisan' orientation towards the emergence of industrialization and professionalization, as described by McKinney[28], though this shift is not always beneficial or desired. Assessment frameworks come with assumptions that sometimes conflict with the reality in many DP situations. Improvement is often oriented towards quality control and consistency, minimizing variability of outcomes over time and reducing individual agency. Culture built around the work of skilled artisans can contrast sharply with these assumptions, resulting in resistance to the transition to an industrial era [28].

The assumptions of sophisticated organizational assessment frameworks such as those compliant with ISO15504[1], a standard for process assessment in software development partially derived from CMM by the SPICE (Software Process Improvement and Capability dEtermination) Working Group[20], include a process orientation, the availability of multiple instances of the assessed processes across the organization's resources, and a depth and distribution of knowledge. These cannot always be assumed. Just as the CMM was not universally praised in the software industry [21], current highly detailed standards prescribing functional requirements for repositories are not necessarily fit for all purposes. Additional tensions of using models that are reflective of 'industrial era' thinking include the tendency to oversimplify reality through CMMs and the obscuring of alternate paths to maturity [44].

The frameworks surveyed here that do explicitly draw on existing CMMs do not distinguish between capability and maturity, project and process, compliance and improvement. Where they do declare adherence to a model such as the SEI CMM, they often do not demonstrate awareness of the concepts and assumptions. In general, greater clarity about underlying concepts and a stronger adherence to design principles for maturity models is needed to instill trust in these frameworks.

Finally, we note that while the CMM approach provides a framework for systematic assessment it focuses on a sequence of events or activities, not on influential factors[44]. Since CMMs were created initially to address process improvement in large organizations devoted to engineering this work is understood to be project-based, and focused on product development. None of these assumptions hold for a typical organization in the domain of DP. DP is often undertaken in small organizations or organizational units, and is not a project-based endeavour resulting in an end product that can be tested for quality and consistency. Some of these assumptions have been dropped in subsequent developments such as CMMI-SVC[47] focused on service delivery; however, current reference models in DP are not based on the principles of service-orientation. Therefore, while addressing capabilities and processes is useful, we may also need to consider the impact of other influential factors over the long-term timeframes necessary for digital preservation.

5.2 Implications for Practitioners

Generally the models surveyed provide limited guidance for conducting an assessment. Together with the absence of empirical evidence, which leads to a lack of trust in the diagnosis, this can present problems for practitioners. Below, we discuss these implications, as well as requirements and expected results, grouping the models by primary purpose.

Certification – There is a clear, but narrow, set of choices for certification: DSA, NESTOR, and ISO16363. These are generally resource-intensive, and make heavy demands on documentation, time, and effort.

The DSA has the least stringent requirements. The process consists of a self-assessment conducted with the online tool and submitted online for review. Required time and resources largely depend on the availability of documentation within the organization. The full self-assessment can take as little as four person days to complete.[14] No site visit from an auditor is required, and the peer review process conducted by the DSA takes approximately two months. Referenced documentation must be made available online for certification. Once granted the seal will need to be updated periodically as the terms of compliance change.

Certification with the NESTOR Seal is similar to the DSA, but has greater demands. NESTOR requires two contact people at the organization to assume responsibility for correspondence during the two-stage review process of the self-assessment and supporting documentation which takes approximately three months.

ISO16363 demands the most of organizations to complete an assessment. The process of certification requires extensive preparation including a thorough self-assessment against the Standard's 84 criteria and the preparation of a full catalog of relevant documentation. This option requires site visits from auditors, who themselves must meet the requirements outlined in ISO16919, and has been shown to take at least six months for many organizations [15].

Choosing from these three options will depend on the particular circumstance of an organization including the availability of documentation, willingness to commit time and resources to the assessment process, and the perceived benefits of certification in relation to the organization's objectives. All three of these assessment frameworks assume a certain degree of maturity and are not oriented towards planning for improvement, but towards compliance with ISO14721 (OAIS). There is potential for a mismatch, if the organization has not adopted the OAIS Reference Model, as David Rosenthal has noted from his experience. [15]

Improvement – As the analysis demonstrated, assessments for improvement vary widely. Decisions in this area can be structured by three factors: (1) the need for tools to conduct the assessment, (2) the major concerns practitioners wish to address, and (3) the availability of third-party assistance.

Nine of the improvement frameworks include various types of tools. Of these, two provide paper templates (CTG and NSLA), three provide both electronic forms and online tools (AIDA and DRAMBORA), and three provide stand-alone online tools (DPCMM, Scoremodel, DSA). Attempting to use these models without reference to the tools provided may compromise the results of the assessment. Third-party online tools may also be available (e.g. for ISO16363).

Organizations seeking third-party assistance in conducting an assessment for improvement have few options. AIDA,

[14] see the Archaeology Data Service (ADS) case http://www.dcc.ac.uk/resources/case-studies/ads-dsa

[15] http://blog.dshr.org/2014/08/trac-audit-lessons.html

CTG, and DPCMM offer different degrees assistance for use of their models. AIDA and CTG are the products of projects and the extent of support and future availability are unknown. Both projects were completed more than five years ago and have shown limited activity in recent years. DPCMM is active through the consulting services, however practitioners should be aware that assessment as a commercial service may have implications for the trustworthiness and reliability of the results. The NSLA model was developed for both internal and external assessment, however it is not clear that third party assistance was ever offered to organizations outside of the consortium.

When selecting a model for organizational assessment, practitioners should be mindful of the fact that with few exceptions, the models for improvement suffer from little or poor documentation, unclear theoretical foundations, and limited transparency. As such, non-certification models raise concerns about reliability and general applicability. All of the improvement models vary on these points, but none are as rigorous as the certification processes.

Initial Planning – Practitioners looking for assessment for initial planning have two options (*Stages* and *Levels*). These models use self-assessment to produce outputs targeted at an internal audience, with less focus on ongoing improvement.

Gaps: Requirements and Outputs – Tensions exist regarding the requirements for, and outputs of, the organizational assessment models currently available in DP. Organizations require well-grounded and robust assessment models with clear methods that produce reliable outputs. The few models that provide full-fledged methods, trustworthy outputs, and meaningful scores also place heavy demands on time and resources that few organizations can afford.

The gap between requirements for organizational assessment in DP and the current range of available options, is particularly significant in light of the degree of development of maturity models in other fields. More sophisticated assessment methods, such as those compliant with ISO15504, make assumptions about process-orientation that do not hold true for many digital repositories seeking assessment. More problematic still, is that many of these models provide ratings that provide the impression of comparability, but without this solid basis. Those gaps point to manifold opportunities for research.

5.3 Implications for Researchers

This survey demonstrates a need for further study of the various types of models available for organizational assessment within DP. First, the work begun here can be extended, and more detailed evaluations of specific models should be completed using principles of maturity models and design science research. Additionally, we have identified the need for further research 'with' models, separated into three areas: development, application, and evaluation of models.

Research Developing Models – Future research can expand on the concept of frameworks and approaches that form the basis of the different models. While the majority of models use a framework related to concepts of capability and/or maturity, few (if any) provide a full definition of these concepts, or demonstrate how they have drawn on the existing research in this area. Shared frameworks provide the benefit of a cumulative tradition, with new work building off the foundations of previous model development. There may be other types of frameworks beyond maturity models that are useful, and an argument can also be made for more diversity in the frameworks used. Additionally, there is a lack of theoretical grounding, or direct evidence of this grounding, in the models studied, particularly around the development of models as designed artifacts. Further work is necessary to determine when and how design principles or guidelines, as described in Design Science research on maturity models, are evident in different models.

Research Applying Models – Research on the application of models is currently limited by the lack of documentation and evidence. Many frameworks do not specify methods (with the exception of DRAMBORA, DSA, and NESTOR), and there is generally little concrete guidance or documentation on carrying out assessments. This is an essential missing component; a robust assessment framework must consist of both a model and a method for its application in order to ensure that assessments are systematic and repeatable. We found that a limited number of case studies exist that describe the details of the application of the model in practice, and these were only available for a handful of models (DSA, ISO16363, and DRAMBORA). This is an area that can and should be explored in greater detail, and in particular there is a need for more rigorous case studies carried out by researchers not associated with development of the model. This is reflected in the recent NDSA Agenda[31] that emphasizes the need for greater large-scale evidence-sharing and capacity-building in the DP community. As noted above, the lack of research on application methods also has implications for practitioners, who may find difficulty in applying the models without clarity of documented methods.

Research Evaluating Models – We found through this survey that, in general, more empirical evidence is needed not only to document applications and report their results, but in order to evaluate and validate the models. While there are case studies available for some models, we did not find that any focus on testing or evaluating the model itself. This is essential to ensure trust in the effectiveness of the assessment framework as an overall tool for improvement. Further evaluation might include more longitudinal studies to identify the critical success factors [43, 45] for DP. As well, engaging evaluation through a Design Science research framework will allow the results to inform the continued iterations and future development of assessment frameworks, models, and methods.

Gaps: Research 'on or about' models – Building on Wendler's distinction of 'research with' maturity models and 'research on or about' maturity models noted in Section 4.1, this survey demonstrates a gap, and need for more 'research on or about' models. This 'meta' approach will benefit the community as it continues to mature.

There is generally limited literature on organizational assessment in DP, and it largely, if not entirely, falls into the category of 'research with models.' Even then, most existing material on organizational assessment in the domain of

DP consists of papers that describe the models, their components and creation. As noted above, more work is needed that studies the development, application and evaluation of models. We propose that drawing connections with Design Science Research can benefit this work, and the field as a whole, by providing a framework to tie all these aspects together, and result in improved models as artifacts for use by organizations.

The lack of work on or about models is not unique to the field of DP, and Wendler concludes that further work is needed to address research on or about models[50]. As others are beginning to address this gap in maturity model research, we can both draw on and align with recent work from other domains. Contributing to this under-represented area will provide the community with a more solid theoretical foundation which will result in better models that are easier to use, more reliable, and more trustworthy. Future work can include concepts and theoretical grounding, definitions and dimensions of maturity, and creation of domain-specific procedures and requirements for maturity model development (such as [11]). We see this paper as a significant contribution, and a starting point for future work in this direction. To the best of our knowledge, this paper is the only contribution to the "meta" field of research on or about maturity models in DP.

5.4 Limitations

We have chosen to undertake a qualitative survey, as opposed to a structured systematic review of literature. It remains an interpretive overview, that has allowed us to characterize many models more generally. Future work could provide a more detailed review of selected models.

6. CONCLUSION & OUTLOOK

This survey attempts to make sense of the diverse and growing landscape of models and tools for organizational assessment and improvement. We have described the options for practitioners seeking to undertake an assessment, and identified trends and gaps for researchers intending to pursue further study of organizational assessment and improvement. Our analysis draws on existing work with maturity models in other domains.

We categorized the models by primary purposes of initial planning, improvement, and certification, and then outlined other requirements for their use. Many options occupy the middle ground between initial planning and more formal certification processes. These range from simple grids to more extensive documents, complete with supporting templates and tools. However, due to lack of empirical evidence, it is still difficult to estimate time and resources required for many of these assessment frameworks, as well as the effectiveness of assessment reports, results and overall organizational outcomes.

Additionally, we have argued here that the concepts of design science research, a growing/emerging approach in information systems, provides effective frameworks for future research and evaluation of models. Future work can benefit from design science principles and guidelines for development of maturity models, that can be adapted to the needs of DP. Design science also, importantly, connects development, application, and evaluation as a cycle, so that application and evaluation continue to inform future development iterations.

We have concluded that further in-depth research and case study evaluation is needed in order to better understand the strengths, weaknesses, and appropriateness of these tools for assessing organizational capabilities. Partnerships and feedback from the community will be essential to undertake this work. We hope to continue this discussion at iPres, to better understand the tensions, needs, and potential synergies of ongoing efforts in the digital preservation community.

Acknowledgements

Part of this work was supported by the Vienna Science and Technology Fund (WWTF) through the project BenchmarkDP (ICT12-046).

7. REFERENCES

[1] ISO 15504:2004 Information technology - Process assessment, 2004.

[2] DIN 31644 Information and documentation - Criteria for trustworthy digital archives, Apr. 2012.

[3] ISO 14721:2012 Space data and information transfer systems - Open archival information system (OAIS) - Reference model, 2012.

[4] ISO 16363:2012 Space data and information transfer systems - Audit and certification of trustworthy digital repositories, 2012.

[5] ISO 16919:2014 Space data and information transfer systems - Requirements for bodies providing audit and certification of candidate trustworthy digital repositories, 2014.

[6] G. Antunes, J. Barateiro, C. Becker, J. Borbinha, D. Proença, and R. Vieira. Project Deliverable SHAMAN Reference Architecture v3.0, 2012.

[7] G. Antunes, J. Barateiro, C. Becker, J. Borbinha, and R. Vieira. Modeling Contextual Concerns in Enterprise Architecture. In *EDOC 2011*, Helsinki, Finland, Sept. 2011.

[8] G. Antunes, D. Proença, J. Barateiro, and C. Becker. Assessing Digital Preservation Capabilities Using a Checklist Assessment Method. In *iPres 2012*, Toronto, ON, Canada, Oct. 2012.

[9] B. Barafort, V. Betry, S. Cortina, M. Picard, M. St Jean, A. Renault, O. Valdés, and P. Tudor. ITSM process assessment supporting ITIL: Using TIPA to Assess and Improve your Processes with ISO15504 and prepare for ISO20000 Certification vol. 217. *Zaltbommel, Netherlands: Van Haren*, 2009.

[10] J. Barateiro, G. Antunes, F. Freitas, and J. Borbinha. Designing Digital Preservation Solutions: A Risk Management-Based Approach. *IJDC*, 5(1):4–17, 2010.

[11] J. Becker, R. Knackstedt, and J. Pöppelbuß. Developing Maturity Models for IT Management - A Procedure Model and its Application. *BISE*, 1(3):213–222, June 2009.

[12] L. Bishoff and E. Rhodes. Planning for Digital Preservation: A Self-Assessment Tool, 2007. NEDCC.

[13] A. Brown. *Practical digital preservation*. Facet Pub., London, 2013.

[14] E. Cardoso. Preliminary results of the survey on Capability Assessment and Improvement. In *iPres 2013*, Lisbon, Portugal, Sept. 2013.

[15] Center for Research Libraries. CRL Report on Scholars Portal Audit, Feb. 2013.

[16] K. Crowston and J. Qin. A capability maturity model for scientific data management: Evidence from the literature. *ASIST*, 48(1):1–9, Jan. 2011.

[17] T. De Bruin, R. Freeze, U. Kaulkarni, and M. Rosemann. Understanding the Main Phases of Developing a Maturity Assessment Model. In B. Campbell, J. Underwood, and D. Bunker, editors, *Faculty of Science and Technology*, pages 8–19, CD-ROM, 2005. AIS, Australasian Chapter.

[18] DigitalPreservationEurope (DPE). Repository Planning Checklist and Guidance DPE-D3.2, Mar. 2008.

[19] C. Dollar and L. Ashley. Digital Preservation Capability Maturity Model (DPCMM) Background and Performance Metrics Version 2.6, May 2014.

[20] A. Dorling. SPICE: Software process improvement and capability Determination. *Information and Software Technology*, 35(6-7):404–406, June 1993.

[21] M. E. Fayad and M. Laitnen. Process assessment considered wasteful. *Comm. of the ACM*, 40(11):125–128, 1997.

[22] A. Goethals. An Example Self-Assessment Using the NDSA Levels of Digital Preservation. In *iPres 2013*, Lisbon, Portugal, Sept. 2013.

[23] A. R. Hevner, S. T. March, J. Park, and S. Ram. Design Science in Information Systems Research. *MIS Quarterly*, 28(1):75–105, Mar. 2004.

[24] T. Jokela, M. Siponen, N. Hirasawa, and J. Earthy. A survey of usability capability maturity models: implications for practice and research. *Behaviour & Information Technology*, 25(3):263–282, May 2006.

[25] A. M. Maier. Assessing Organizational Capabilities: Reviewing and Guiding the Development of Maturity Grids. *IEEE Transactions on Engineering Management*, 59(1):138–159, Feb. 2012.

[26] N. Y. McGovern and A. R. Kenney. The Five Organizational Stages of Digital Preservation. In P. Hodges, M. Bonn, M. Sandler, and J. P. Wilkin, editors, *Digital Libraries: A Vision for the 21st Century*. Michigan Publishing, University of Michigan Library, 2003.

[27] A. McHugh, R. Ruusalepp, H. Hofman, et al. Digital Repository Audit Method Based on Risk Assessment (DRAMBORA), 2007. `eprints.erpanet.org/122/`.

[28] P. McKinney. From Hobbyist to Industrialist. Challenging the DP Community, Oct. 2012. Open Research Challenges Workshop at *iPres* 2012, `digitalpreservationchallenges.wordpress.com`.

[29] T. Mettler and P. Rohner. Situational Maturity Models as Instrumental Artifacts for Organizational Design. In *DESRIST '09*, Malvern, PA, May 2009.

[30] National Library of Australia. UNESCO Guidelines for the preservation of digital heritage, 2003.

[31] NDSA. 2015 National Agenda for Digital Stewardship, Sept. 2014.

[32] NESTOR Working Group. Catalogue of Criteria for Trusted Digital Repositories, Dec. 2006.

[33] NESTOR Working Group. Explanatory notes on the nestor Seal for Trustworthy Digital Archives, 2013.

[34] OCLC, CRL. Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist, Feb. 2007. Ver 1.0.

[35] M. C. Paulk, B. Curtis, M. B. Chrissis, and C. V. Weber. Capability Maturity Model for Software, Version 1.1, Feb. 1993. CMU/SEI-93-TR-024.

[36] D. Pearson and L. Coufal. Digital Preservation Environment Maturity Matrix, Nov. 2013. National and State Libraries of Australasia, `www.nsla.org.au`.

[37] E. Pinsent. The AIDA self-assessment toolkit Mark II, Feb. 2009. ULCC, `http://aida.jiscinvolve.org`.

[38] J. Pöppelbuß and M. Röglinger. What makes a useful maturity model? A framework of general design principles for maturity models and its demonstration in business process management. In *ECIS*, 2011.

[39] T. A. Prado, A. M. Cresswell, S. S. Dawes, B. Burke, L. Dadayan, S. Embar, and H. Kwon. Building State Government Digital Preservation Partnerships: A Capability Assessment and Planning Toolkit, Version 1.0, 2005.

[40] D. Preença, R. Vieira, and J. Borbinha. Project Deliverable D7.2 E-ARK - Archiving Maturity Model - Initial Assessment and Evaluation, Oct. 2015. Project 620998: European Archival Records and Knowledge Preservation, `www.eark-project.com`.

[41] D. Proença, R. Vieira, and J. Borbinha. Project Deliverable D7.1 E-ARK - A Maturity Model for Information Governance - Initial Version, Feb. 2015. Project 620998: European Archival Records and Knowledge Preservation, `www.eark-project.com`.

[42] Research Libraries Group. Trusted Digital Repositories: Attributes and Responsibilities, May 2002.

[43] J. F. Rockart. Chief executives define their own data needs. *Harvard Business Review*, 57(2):81–93, 1979.

[44] M. Röglinger, J. Pöppelbuß, and J. Becker. Maturity models in business process management. *BPMJ*, 18(2):328–346, Apr. 2012.

[45] M. Rosemann and T. De Bruin. Application of a Holistic Model for Determing BPM Maturity. *BPTrends*, Feb. 2005.

[46] P. Sinclair, J. Duckworth, L. Jardine, A. Keen, R. Sharpe, C. Billenness, A. Farquhar, and J. Humphreys. Are you Ready? Assessing Whether Organisations are Prepared for Digital Preservation. *IJDC*, 6(1):268–281, Nov. 2011.

[47] Software Engineering Institute, CMU. CMMI for Services, v1.3, Nov. 2010. CMU/SEI-2010-TR-034.

[48] Software Engineering Institute, CMU. Appraisal requirements for CMMI (ARC), v1.3, 2011. CMU/SEI-2011-TR-006.

[49] S. Vermaaten, B. Lavoie, and P. Caplan. Identifying threats to successful digital preservation: The SPOT model for risk assessment. *D-Lib*, 18(9):4, 2012.

[50] R. Wendler. The maturity of maturity model research: A systematic mapping study. *Information and Software Technology*, 54(12):1317–1339, Dec. 2012.

DataNet Federation Consortium Preservation Policy ToolKit

Reagan Moore
University of North Carolina at Chapel Hill
312 Lenoir Dr
Chapel Hill, NC 27599
919 962 9548
rwmoore@renci.org

Arcot Rajasekar
University of North Carolina at Chapel Hill
312 Lenoir Dr
Chapel Hill, NC 27599
919 966 3611
sekar@renci.org

Hao Xu
University of North Carolina at Chapel Hill
312 Lenoir Dr
Chapel Hill, NC 27599
919 962 9548
xuh@cs.unc.edu

ABSTRACT

The DataNet Federation Consortium uses a policy-based data management system to apply and enforce preservation requirements. This paper describes the Preservation Policy Toolkit developed by the consortium. In particular, the paper describes the infrastructure needed for preservation, presents examples of computer actionable forms of policies, and provides a generic template for designing actionable preservation policies.

General Terms

Preservation strategies and workflows.

Keywords

Policy-based data management, preservation policies, computer actionable procedures.

1. INTRODUCTION

The NSF DataNet Federation Consortium (DFC) infrastructure enables multiple Science and Engineering communities to implement their preferred data management applications and establish trusted research collaborations [1]. Partners within the DFC have implemented a variety of data-centric environments including data preservation systems (archives), data sharing systems, data publication systems (digital libraries), data distribution systems, and data processing systems (processing pipelines) to serve the needs of their specific communities and research groups. The DFC accommodates each type of data management application by specifying a set of policies that enforce the desired properties for that type of data management application:

- A trusted digital archive focuses on properties related to: authenticity; integrity; access control; chain of custody; persistent storage; fidelity; and original arrangement.
- A data sharing environment focuses on properties related to: unified name spaces for users, files, and collections; metadata-based discovery; access controls; auditing; hierarchical arrangement; and ease of access.
- A digital library focuses on: controlled name spaces for files, collections and metadata; descriptive metadata standards; standard data formats; multi-faceted search; and logical collection arrangements.
- A data distribution system focuses on: fault tolerance; automatic failover; on-demand caching; replication; synchronization; staleness control; high availability; streaming; and high-speed content delivery.
- A processing pipeline focuses on: controlled name spaces for users, files, collections, and procedures; distributed service and workflow automation; cloud computing; scheduling of high-performance computation; third-party and licensed service invocation; workflow reuse; repurposing of workflows; and provenance of workflows.

Each of these types of data management applications can build upon generic data grid infrastructure by choosing an appropriate set of policies and procedures. The DFC uses the integrated Rule Oriented Data System (iRODS) data grid software [2] as a platform to implement community-specific management policies. The policies determine when and where procedures are executed. Policies can be automatically enforced at policy enforcement points that are encoded in the software middleware within the iRODS system, or policies can be executed interactively by a user or grid administrator, or policies can be scheduled for deferred and periodic execution. The policy enforcement points typically control management policies. Deferred and periodic execution is used for administrative tasks. Interactive execution is used by users to launch remote workflows and is also used to validate assessment criteria.

The DFC is developing toolkits for each of the data management applications outlined above. This paper describes the Preservation Policy Toolkit (PPTK). The PPTK can be tuned, modified or extended by each Science and Engineering community to meet their particular requirements. In the next section, we describe the concepts behind the implementation of policies within iRODS, followed by a discussion of policy templates and policy languages, and summarize the elements in the PPTK. Several examples of policies are provided as part of the discussion.

2. POLICY CONCEPTS IN DFC

In this paper, we discuss the preservation environment needed to implement data management applications such as a trusted digital archive that automates policy enforcement within cyber-infrastructure. A preservation environment can be defined by the set of policies and procedures that enforce the properties of authenticity, integrity, access control, chain of custody, persistent storage, fidelity, and original arrangement. In Figure 1, a generalization of this approach for implementing preservation properties is shown. Given a specific preservation purpose, an archive can be assembled that has desired properties such as integrity enforcement, arrangement, and access controls. The properties themselves may have associated requirements such as completeness (all files in the archive have the same property), correctness (incorrect values for information properties have been identified and isolated or eliminated), consensus (the properties represent the combined desire of the group assembling the archive), and consistency maintenance (the same metadata and data format standards have been applied to all files in the archive). Each desired property is enforced by a set of policies that determine when and where associated procedures are executed.

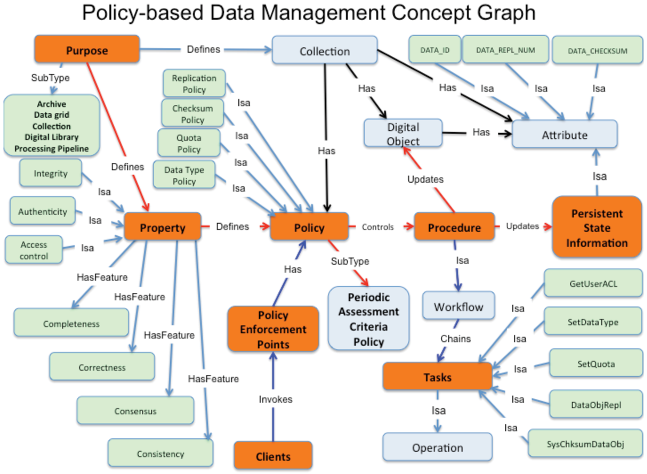


Figure 1. Policy Concepts

The associated procedures are implemented as workflows constructed by chaining basic tasks or functions (called micro-services) that are provided in the iRODS data grid. The functions implement basic operations such as generate a checksum, or replicate a file, or set the data type. The results of applying the functions are saved as persistent state information or metadata attributes on the name spaces for files, collections, users, storage systems, metadata, policies, and micro-services.

Consider the integrity maintenance property. In an implementation, one may perform such integrity maintenance by applying policies for generation of checksums and replication of files. A checksum is used as a digital signature to verify the fidelity and integrity of the deposited material in the archive. In some rigorous applications, more than one type of digital signature (using different algorithms) may need to be maintained as part of the digital collection. Replication is enforced to recover from disasters and failures. Periodic verification of checksums together with management of replicas provides a means to identify file corruption and rectify through synchronization with a high fidelity copy. Policies are needed to set the number of required replicas, set the verification periodicity, and define the mode of failure recovery. Additional policies apply this state information to enforce the integrity property when files are ingested into the archive.

In essence, policy-based preservation systems encapsulate four foundational concepts:

1. Purpose for creating the preservation archive expressed as the management of a set of desired properties.
2. Consensus on preservation enforcement as a set of desired policies.
3. Maintenance of preservation properties through a set of required procedures.
4. Tracking of preservation state information through required attributes assigned to the controlled name spaces for users, files, collections, and micro-services.

3. POLICY TEMPLATES

This view of preservation as the set of properties that will be maintained over time is consistent with the ISO 16363 standard [3]. Each of the trustworthiness metrics expressed by the standard can be captured in policies that are automatically enforced by the data management environment. The PTAB ISO 16363 Metric Knowledge Base lists a set of required supporting evidence for each metric. For example:

> “**4.6.1 The repository shall comply with Access Policies.**
> 1. Access policy for repository.
> 2. Collection Development Policy.
> 3. Definition of the Designated Community.
> 4. Demonstrations and discussion with relevant staff of what occurs when a query results in 'Access Denied'.
> 5. Documentation that illustrates the Access Policy is being carried out: Sign in sheets, logs of access, logs of successful and unsuccessful access to the system, follow up emails or help desk reports when 'access denials' received.
> 6. Examples of Preservation Description Information (PDI) that contain Access Rights information.
> 7. If there are access controls on private or restricted content, then particular events when the content was accessed by users or staff should be checked.
> 8. License agreements for content.
> 9. Mission Statement.
> 10. Relevant Copyright law.
> 11. Submission agreement(s).
> 12. User surveys or interviews that determine user satisfaction with delivery of DIP’s.”

Policies can be implemented through procedures that generate the required evidence for each metric. Policies can be created that identify the location of the required documentation, generate event information for each action, and log the results of all access checks. These policies can be implemented as machine-actionable procedures, enabling automation of preservation tasks.

A template that captures the information associated with a policy has been published by the Research Data Alliance Practical Policy working group [4]. In Table 1, an example is provided for specifying policies for access controls.

The example has two sections: the set of state attributes needed to decide when to execute the policy, and the set of state attributes needed while executing the policy.

The template can be used to design the set of controlling policies and execution procedures that implement the evidence specified for each ISO 16363 metric. The template lists the policy name, the constraints that limit application of the policy, the state information needed to evaluate the constraints, the operations that the policy will apply, and the persistent state information that is needed or changed by the policy.

The constraints imposed on the policy define how the policy should be applied. In this case, the archive may choose to enforce access controls by the role assigned to each user (administrator, user) or by a unique identifier for each user (account name). The access controls may be applied at the collection level or at the individual file level. Choosing the type of access control to implement defines the state information that will be needed.

The operations performed for controlling access include:

- Creating identifiers for persons, collections, and files.
- Assigning roles to persons.
- Assigning access controls to collections and files (in effect a relationship between the person identifier and the file identifier).
- Assigning inheritance of access controls on collections (files can inherit the access control of the collection).
- Checking access permissions on reads and for other actions on the file.
- Verifying the set of access controls applied to files in a collection.

Table 1. Policy Template for Access Control

Policy type	**Constraint**	**State attributes for Constraint**
Access data	By role (type of person)	User_ID
		Role_type per User_ID
		Role_ACL
	By ACL (read permission)	User_ID
		File_name
		ACL per File_name per User_ID

	Operations	**State Attributes for Operation**
	Set person name	User_ID
		User_name
	Set file name	File_ID
		File_name
	Set role per person	User_ID
		Role_type
	Set ACL on file	File_ID
		User_ID
		ACL_type
	Set sticky bit on collection	Collection_name
		Sticky-bit_value
	Set access on replication	File_ID
		Replica_number
		User_ID
		ACL_type
	Execution - check ACL on read	File_name
		User_ID
		ACL_type
	Verify ACLs	File_ID
		Replica_number
		User_ID
		ACL_type

One can immediately notice that the evidence listed for the access control metric in ISO 16363 needs to be augmented with policies that are driven by the type of implementation. The preservation environment has to map from the metric evidence specification to the technologies that are currently available for implementation of the archive. Depending upon the choice of technology, different mappings will be required. For example:

- Choice of person identifier depends upon the type of authentication system that is used (certificate authority, LDAP directory, one-time password, ORCID).
- Choice of file identifier depends upon the type of storage system (Unix file system, tape archive, object store) and the object identifier (GUID, OID, handle, logical name).
- Choice of role-based or account-based access controls depends on the type of user authentication environment.
- Choice for identification of copies of files (replicas, backups, versions) depends upon the required persistence properties.

A second observation is that the documents specified in the audit checklist can be supported by generic policies. Thus policies for storing, finding, and retrieving documents can be used to archive the collection development policy, the definition of the designated community, examples of preservation description information (PDI) that contain access rights information, license agreements for content, mission statement, relevant copyright law, submission agreement(s), and user surveys or interviews that determine user satisfaction with delivery of DIP's. The document attributes may need to be organized and associated with either a user name space, or a collection name space, or individual files.

A third observation is that sign-in sheets, logs of access, logs of successful and unsuccessful access to the system, and follow up emails or help desk reports when 'access denials' are received can be supported by generic event management policies. If the archive is able to encapsulate information about all actions that are performed in standard events, then the events can be saved and indexed. A generalization of this is the ability to map from:

- An action that was taken (record ingestion, user access, archive administrator process),
- To the operation that was performed within the archive,
- To the state information change that resulted from the action.

It should then be possible to identify all interactions with the archive and verify that the resulting operations were consistently applied. This includes application of access controls, or maintenance of file integrity, or creation of AIPS, or tracking of submissions. An audit trail can be saved as the sequence of events that changed the archive state information. The events can be indexed and analyzed for compliance with the desired archive properties. In addition, all changes to the preservation environment state information can be correlated with a controlling policy.

In summary, multiple types of policies may be needed for each type of evidence:

1. Policies to set input parameters (environmental variables) needed for policy execution.
2. Policies to control execution of a procedure.
3. Policies to automate execution of administrative functions, typically performed by the archive administrator.
4. Policies to verify compliance with the desired preservation properties.

The policies may be run interactively by the archive administrator (policy type 1), or enforced at a policy enforcement point within the software (policy type 2), or executed periodically by the rule engine (policy types 3 & 4). The policies are organized into a preservation policy toolkit. Each community that requires preservation can modify the policy toolkit to implement their required preservation policies.

4. POLICY VIRTUALIZATION

The choices made today for implementing an archive will change as better technology emerges. This raises an immediate challenge for preservation environments. How can the same policies be effectively applied in the future? How can the effort to migrate to new technologies be minimized? How can federation across multiple archive implementations be achieved? Note that migration to new technology and federation across heterogeneous technologies are effectively the same capability. At the point in time when new technology is acquired, both the old technology and the new technology will be present in the system. Records can be migrated from the old technology to the new technology using federation mechanisms. The ability to federate across technology implementations is essential for continued enforcement of preservation policies over time.

Policy-based data management systems such as data grids handle technology evolution through use of virtualization mechanisms. Interactions with technology are done through software middleware that map from the desired action to the protocol required by the technology choice. The software that does the mapping is encapsulated in a pluggable driver, enabling the replacement of the old technology by plugging in a driver for the new technology. Pluggable drivers are used within the DFC for interactions with authentication systems, storage systems, databases, network transport, rule engines, and micro-services (basic operations). Through plugins, a preservation environment can interact with multiple types of systems simultaneously, and manage migration to new technologies.

Virtualization also implements the ability to manage all of the properties of a preservation environment independently of the choice of technology. This includes management of the names of the users, the names of the files, the organization of files into collections, the provenance and descriptive metadata, the access controls, and administrative metadata such as checksums, file size and storage location. The information is stored as metadata in a database.

For example, consider the addition of a file to the system. Even though the explicit event is a simple file addition, the response of the system may require the execution of multiple policies, with each policy potentially executing procedures that manipulate multiple types of objects. Policies that are executed may include:

- Authentication of the person adding the file.
- Authorization for the addition of a file.
- Evaluation of a storage quota for the storage resource.
- Creation of a persistent identifier for the file.
- Validation of the Submission Information Package against the submission agreement.
- Logical arrangement of the file as a member of a collection (creation of a logical file name).
- Selection of a storage resource for the physical copy of the file.
- Creation of a physical file name on the storage resource
- Inheritance of access controls from the collection access controls.
- Creation of a checksum.
- Creation of a persistent object (storage of the file as received).
- Replication of the persistent object to a second storage location.
- Assignment of a retention period for the file.
- Assignment of a disposition procedure to the file.
- Assignment of a data type to the file based on the file extension.
- Creation of a copy with a required data format.
- Storage of system level metadata (owner name, access controls, checksum, file size, replica location, retention period, file type).
- Extraction and storage of descriptive metadata.
- Creation of an Archival Information Package (aggregation of metadata with the file into a container).
- Storage of the AIP.
- Replication of the AIP.
- Generation of event information for each step of the ingestion.
- Storage and indexing of the event information.

Policies can be defined that control each of the ingestion steps. It is then possible to associate different ingestion steps with different collections. Also the policies may need to evolve over time to handle changes in technologies, or changes in management, or changes in preservation standards. This will require support for multiple versions of policies, with different sets of constraints applied within each version. The policies will need to be archived along with the records to enable a future archivist to track how each record was controlled over time. It should be possible for a future archivist to start with an original Submission Information Package, apply the sequence of policies recorded in event information, and re-generate the current Archival Information Package.

4.1 State Information

Virtualization depends upon having a "complete" set of state information (metadata attributes) that can be queried and retrieved. Information is needed about the preservation environment for each step in the file ingestion process. This includes information about not only each record (representation information, provenance information, description information), but also information about the preservation environment (user names, storage locations, policies).

Typical file system state information is listed in Table 2. The information stored about each file is quite limited. A preservation environment augments this information with provenance information, representation information, description information,

Table 2. File System State Information

File Name
File Location on disk
Creation time
Modification time
File size
Access control
Locks
Soft Link
Directory

and administration information. In practice, the DFC manages more than 330 state information attributes about both the records and the preservation environment. Information is managed about users, files, collections, storage resources, metadata, rules, micro-services, quotas, system load, audit trails, and federations.

4.2 Operations

Virtualization depends upon having a complete description of all the operations that will be performed within the preservation environment. The operations performed upon a file system typically consist of create, open, close, read, write, update, seek, stat, chown, link, and unlink. Some of the operations may be applied to a file or to a group of files. Preservation environments require support for additional operations such as creation of checksums, replication, migration, and format transformation.

A generic characterization of operations performed within data management systems is needed. To base the discussion on well-known concepts, consider the characterization of file systems shown in Figure 2. The file system comprises an environment that is defined by the state information maintained about each file. Interactions with the file system consist of events that specify an operation. Each operation manipulates a file and changes the associated state information. Operations may require access to state information such as file location, or file size, or file owner.

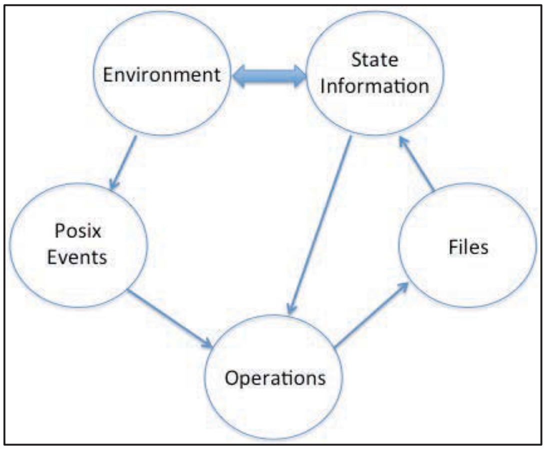


Figure 2. File System Characterization

Interactions with the files are done through interactive execution of clients, which invoke the desired operation through a system call. This approach makes it possible to implement a standard data management interface on different types of hardware systems, which in turn enables the migration of files across storage systems.

We can generalize this model of data management by introducing policies that control the operations performed within the system. In Figure 3, we introduce three significant changes:

1) Operations are replaced by policies.
2) Files are replaced by objects.
3) Updates on objects and on state information are implemented as procedures.

Additional operations can be added to the system through the creation of new procedures. The knowledge needed to manage the procedures can be captured in policies, and the information needed to execute the new procedures can be added as additional metadata. This makes it possible to add operations to the preservation environment, along with the new policies and state information. The preservation environment can now evolve to track changes in preservation requirements, changes in technology, and changes in administration.

Within the DFC, procedures are implemented as workflows that are created by composing together basic functions, called micro-services. The DFC supports more than 300 micro-services that implement data management operations. The micro-services can be categorized as operations for user management, file manipulation, collection management, metadata manipulation, policy management, network management, messaging, administration (setting environment variables, quotas, load monitoring), and data grid manipulation (federation).

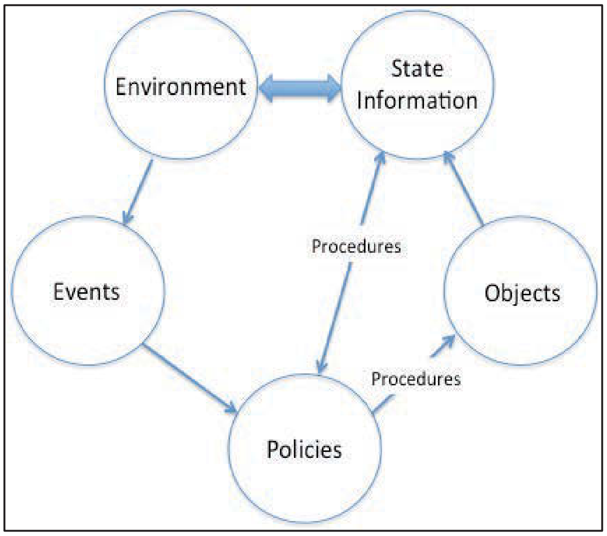


Figure 3. Policy-based Data Management

Micro-services can be created that support interaction with specific types of technology. A typical example is the creation of a micro-service that supports access to a remote service for file conversion. The micro-service manages the interaction with the network protocol required for communication with the remote service. Since multiple types of technology exist today, this requires support for versions of micro-services, as well as versions of state information.

5. POLICY LANGUAGE

Policies within the DFC are implemented as workflows, by chaining together micro-services. A rule engine is used to parse each workflow, evaluate the policy constraints, invoke execution of each micro-service, and manage errors. The workflow language had to be Turing complete, enabling the creation of workflows that included conditional tests and loop constructs. A typical policy would specify a constraint as a conditional test on system state information and session variables, generate a query that is sent through a catalog interface to a database, loop over the database query results, apply arithmetic operations and string manipulation to variables, and write results to standard out for interactive execution or to a file for storage within the data grid. In the DFC, new policies can be added dynamically to the system through inclusion in a rule base.

The choice of where and when to apply the policies is mediated through the use of policy enforcement points within the data grid software middleware. In the DFC, the locations of the original policy enforcement points were hard-coded. Through extensions developed by the iRODS Consortium [5], policy enforcement points were made pluggable. Each time a new operation is added, pre-process and post-process policy enforcement points are added automatically.

A consequence of the micro-service plugin extension is that now every operation performed within the preservation environment can be tracked, along with the corresponding change to state information. The state changes can be saved as events that are indexed in an external indexing system. The events can be analyzed to verify compliance over time with the desired properties of the preservation environment.

The design of preservation policies that will be executable in the future is based on the assumption that the knowledge needed to interact with technology can be encapsulated in versions of micro-services. By invoking the current micro-service version, a policy will remain executable. This in turn requires that the preservation environment manage all information needed to apply the procedure within a metadata catalog, independently of the choice of storage technology. The catalog can then be queried to retrieve the information needed to apply the current micro-service version.

The policy language is interpreted by the rule engine. To enable long-term preservation, the rule engine itself had to be pluggable, enabling the use of a new rule engine and a new rule language by future archivists. The DFC preservation environment thus provides multiple levels of virtualization:

- From actions requested by clients to standard operations supported by the data grid.
- From the state information maintained by the data grid to the information required by the selected storage technology.
- From the knowledge encapsulated in micro-services to the execution of the standard data grid operations.
- From the standard operations supported by the data grid to the operations provided by the selected storage technology.
- From a consensus on management decisions to choice of policies enforced at policy enforcement points within the data grid.

With these levels of virtualization, a preservation environment can be created that is technology independent, enabling the incorporation of new technologies over time while maintaining persistent objects.

An example of the rule language is shown in Figure 5. Each workflow operation (variable assignment, string concatenation, foreach loop, conditional if test) is treated as a micro-service. The rule engine parses each line in the workflow, invokes the associated micro-service, and manages information exchanges between micro-services through in-memory data structures. The workflows can be distributed across multiple servers. Information exchange between servers is mediated by packing instructions that serialize the in-memory data structures, send the result over the network to the next participating server, and unpack the information into a local in-memory data structure in the remote system.

Policies are stored at each server in a distributed rule base. This improves performance, makes it possible to distribute the policy enforcement across all participating storage resources, and makes it possible to install different policy sets at each server. One consequence is that a distributed debugger is needed to analyze problems in distributed workflows. This capability is provided within the DFC infrastructure through use of a messaging system.

6. PRESERVATION POLICY TOOLKIT

The DFC has developed a set of policies required for preservation. The policies (forming a toolkit) are driven by community requirements and represent instances of computer actionable rules that control administrative operations. The policies are driven by local security requirements, local storage facilities, local authentication requirements, and local networking infrastructure. The examples provided in this paper are intended to illustrate some of the challenges in writing computer actionable rules. The rules are modifiable for application in other preservation environments.

6.1 Sample Policy: Network Firewall

Implementing policies for a preservation environment is a complex task. A standard challenge in implementing a preservation environment is management of network firewalls. If an archive storage resource is located behind a firewall, prohibiting access from external networks, then policies are needed to manage ingestion. One approach is to implement data staging, with records deposited into a network accessible storage system as shown in Figure 4.

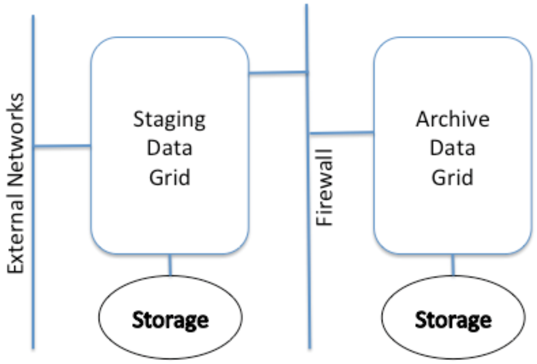


Figure 4. Deep Archive

A policy running within the Staging Data Grid analyzes the Submission Information Packages for compliance with a submission agreement, checks for the presence of viruses, and sets an approval flag for qualified data. A policy that runs within the Archive Data Grid queries the external Staging Data Grid and pulls the approved files into the archive.

A version of the staging policy that implements multiple operational steps needed for a production environment is shown in Figure 5. Files are copied from the staging area into an archive by a policy running on the staging area. The rule implements the following steps:

- Use session variables to find the data grid and account name under which files will be accessed. $rodsZoneClient is the name of the staging data grid. $userNameClient is the account name on the staging data grid.
- Create path names for the staging directory and the archive directory.
- Get the current system time in a human readable format.
- Check whether a directory exists in the archive for storing log files.
- Create the directory if needed.
- If the log directory cannot be created, fail with an error message.
- Create the log file for tracking data storage operations.
- Create a query to list the files and their checksums in the staging data grid.
- Execute the query and loop over the result set.

- Extract each file name and checksum value.
- Copy each file to the archive and force an overwrite of existing files.
- Set ownership access controls on each file.
- Calculate the checksum of each file after it is moved.
- Verify the checksum is correct.
- For files moved successfully, delete the copy in the staging area.

Variants of the rule are used to execute the rule from the archive and pull data from the staging area, initiate the original archive, and push data to a second archive.

```
myStagingRule {
# Loop over files in a staging area,
#/$rodsZoneClient/home/$userNameClient/*stage
# Put all files into collection
#/*DestZone/home/$userNameClient#$rodsZoneClient/*Coll

  *Src = "/$rodsZoneClient/home/$userNameClient/*Stage";
  *Dest= "/*DestZone/home/$userNameClient"
++"#$rodsZoneClient/" ++ *Coll;

#=get current time, Timestamp is YYY-MM-DD.hh:mm:ss  =====
  msiGetSystemTime(*TimeH, "human");

#=create a collection for log files if it does not exist ===========
  *LPath = "*Dest/log";
  *Query0 = select count(COLL_ID) where COLL_NAME = '*LPath';
  foreach(*Row0 in *Query0) {*Result = *Row0.COLL_ID;}
  if(*Result == "0" ) {
    msiCollCreate(*LPath, "0", *Status);
    if(*Status < 0) {
      writeLine("serverlog", "Could not create log collection");
      fail;
    }  # end of check on status
  }  # end of log collection creation

#= create file into which results will be written ===============
  *Lfile = "*LPath/Check-*TimeH";
  *Dfile = "destRescName=*Res++++forceFlag=";
  msiDataObjCreate(*Lfile, *Dfile, *L_FD);
#============ find files to stage ======================
  *Query = select DATA_NAME, DATA_CHECKSUM where
COLL_NAME = '*Src';
  foreach(*Row in *Query) {
    *File = *Row.DATA_NAME;
    *Check = *Row.DATA_CHECKSUM;
    *Src1 = *Src ++ "/" ++ *File;
    *Dest1 = *Dest ++ "/" ++ *File;
# ============Move file and set access permission =========

msiDataObjCopy(*Src1,*Dest1,"destRescName=*Res++++forceFlag=
", *Status);
    msiSetACL("default", "own", $userNameClient, *Dest1);
    writeLine("*Lfile", "Moved file *Src1 to *Dest1");
# =========== verify checksum ==========================
    msiDataObjChksum(*Dest1, "forceChksum=", *Chksum);
    if(*Check != *Chksum) {
      writeLine("*Lfile", "Checksum failed on *Dest1");
    }
# =====  Delete file from staging area if checksum is good ========
    else {
      msiDataObjUnlink("objPath=*Src1++++forceFlag=", *Status);
    }
  }
}
INPUT *Stage =$"stage", *Coll=$"Archive",
*DestZone=$"tempZone", *Res=$"demoResc"
OUTPUT ruleExecOut
```

Figure 5. Staging Policy

Additional policies are needed to check access controls on files in the archive, verify checksums periodically, verify presence of required metadata, and identify file types. For policies that have been verified to work correctly, the execution of the rule can be automated. Rules can be run periodically, or executed at policy enforcement points. The choice usually depends upon whether batch processing is preferred, or whether continuous processing is needed to manage the workload. The archive administrator has control over the policies that are being applied.

The DFC preservation policy toolkit contains multiple policies for preservation, which can be categorized at an abstract level as:

- Authenticity
- Integrity
- Authorization
- Chain of custody
- Persistent storage management
- Ingestion
- Dissemination
- Fidelity
- Original arrangement and
- Packaging

As was shown with the firewall maintenance policy (which is central to ingestion and dissemination) similar integration of operational procedures had to be done for other abstract policies. In many instances, the technology needed to apply the rule is implemented in an external system. The systems identified in parentheses within the following preservation task list show external services that have been integrated into the DFC environment for providing preservation functionalities. The toolkit contains several snippets of code that can be chained to enable creation of additional policies:

- Automate application of access restrictions.
- Transform data sets to non-proprietary formats.
- Generate event preservation metadata.
- Automate enforcement of user submission agreements.
- Automate creation of checksums.
- Automate capture of description metadata.
- Automate data archiving.
- Automate de-identification of data sets (BitCurator [6]).
- Apply unique identifiers to data (Handle system [7]).
- Enforce authentication of users (InCommon [8]).
- Map metadata terms across ontologies (HIVE [9]).
- Export data in multiple formats (NCSA Polyglot [10]).
- Track usage (Databook).
- Check for viruses (ClamScan [11]).
- Control data retention period.
- Control data disposition.
- Control searches.
- Generate storage cost reports.
- Replicate datasets.
- Copy datasets.
- Synchronize datasets.
- Verify checksum.
- Verify metadata compliance.
- Verify access control against requirements.
- Verify arrangement against requirements.
- Verify format compliance (e.g. XML).

7. COMPARISON WITH ISO 16363

The viability of the DFC preservation approach can be evaluated through comparison with prior preservation audit checklists. Specifically, can each of the tasks defined in prior checklists be turned into computer actionable rules?

An analysis of the ISO 16363 audit checklist has been done to identify which tasks can be automated. The analysis identified 140 preservation tasks. By casting the tasks in terms of generic operations, the number of tasks can be minimized. This requires identifying the state information that will be needed when applying the generic task. An example is a generic rule to print a report. The required state information is the location of the report (logical name) within the preservation environment.

For each task, the predominate operation has been identified, along with the type of entity that is being manipulated. Seven generic operations were defined:

Create, Read, Update, Delete, Copy, Move, & Execute.

The operations were applied to seven object types:

File, Metadata, Events, Policies, Procedures, Database & Ontology.

Examples of the operations upon objects are shown in Table 3. A representative task is selected for inclusion in the list for each combination of operation and object. Thus the "Create" operation can be applied to files, metadata, policies, procedures, events, databases and ontologies. Each task actually may involve multiple operations. Thus an integrity check will verify checksums, delete bad copies, and replace the bad copies from a good replica.

Table 3. Computer actionable task list for ISO 16363

Operation	Object	Task
Copy	file	Create authentic copy from master, verify checksums
Create	Database	New database from metadata in a federated archive
Create	events	Record all micro-services applied to file, along with state information
Create	file	Generate AIP based on AIP template
Create	metadata	Create GUID, handle and logical name for record
Create	ontology	Ontology for designated community terms
Create	policies	Set access policies from remote federation
Create	procedure	Create queries on descriptive metadata
Execute	procedure	Apply transformative migration on format
Move	file	Migrate records to new storage resource
Read	events	List persons who applied archival functions, or accessed file
Read	files	Verify presence of all records specified in submission agreements
Read	metadata	List all persons with access to a collection
Read	policies	List rules for collection
Read	procedure	Verify mechanisms for mitigating risk of data loss
Update	ontology	Remove obsolete terms, incorporate new terms

A second observation is that multiple tasks were required for each criterion specified in the ISO 16363 audit checklist. This raises the question for whether it is possible to identify fundamental criteria that reduce a task to a single operation on a single type of object. Based on this analysis, this will be very difficult to do, since each criterion currently accesses multiple state information attributes to correctly apply a generic operation, interacts with multiple file replicas, and generates multiple event notifications.

The objective of creating computer actionable policies for each task remains a viable approach to preservation. Generic operations can simplify the implementation of preservation tasks while policies can manipulate the multiple objects needed to execute the preservation tasks. This makes it possible to automate preservation processes.

8. SUMMARY

Policy-based data management systems enable creation of preservation environments that maintain records in their original form (persistent objects), while managing interactions with the changing technology in the external world. A preservation environment enables:

- Communication with the future. Records archived today can be retrieved by a future archivist.
- Validation of communication from the past. An archivist can verify the set of policies that governed preservation of a record.
- Management of new technology. A preservation environment allows the flow of technology through the archives while preserving the original records. As new technology becomes available, the technology can be incorporated into the archive without affecting the persistent objects.

Policies are used to enforce assertions that are made about the properties of the preservation environment. Policies are periodically executed to verify the assertions, since storage systems may fail, networks may fail, operators may run obsolete procedures, and software system may malfunction. All assertions made about a preservation environment have to be verified over time. Automating validation of assessment criteria is essential when making assertions such as trustworthiness of a repository.

A generic policy template can be used to define the required policy components. Based on the policy toolkits developed within the DFC, a generic policy template includes:

- Policy name,
- Constraints controlling policy application,
- State information evaluated by the constraints,
- Operations performed by the policy,
- State information needed for operation execution.

With this information, policies can be implemented that automate each preservation task.

9. ACKNOWLEDGMENTS

The development of the iRODS data grid was funded by the NSF OCI-1032732 grant, "SDCI Data Improvement: Improvement and Sustainability of iRODS Data Grid Software for Multi-Disciplinary Community Driven Application," (2010-2013). The results presented in this paper were funded by the NSF Cooperative Agreement OCI-094084, "DataNet Federation Consortium", (2011-2015). The iRODS Consortium developed the pluggable architecture used by the DFC.

10. REFERENCES

[1] Moore, R., A. Rajasekar, "Reproducible Research within the DataNet Federation Consortium", International

Environmental Modeling and Software Society 7th International Congress on Environmental Modeling and Software, San Diego, California, June 2014, http://www.iemss.org/society/index.php/iemss-2014-proceedings.

[2] Rajasekar, A., Wan, M., Moore, R., Schroeder, W., “Micro-Services: A Service-Oriented Paradigm for Scalable, Distributed Data Management”, in “Data Intensive Distributed Computing”, January 2012, ISBN13: 9781615209712, pp. 74-93.

[3] ISO 16363:2012, “Space data and information transfer systems – Audit and certification of trustworthy digital repositories”, http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510.

[4] Moore, R., R. Stotzka, C. Cacciari, P. Benedikt, “Practical Policy Templates”, submitted to Research Data Alliance as a deliverable of the Practical Policy Working Group, February, 2015.

[5] iRODS Consortium, http://irods.org.

[6] Lee, Christopher A., Kam Woods, Matthew Kirschenbaum, and Alexandra Chassanoff. “From Bitstreams to Heritage: Putting Digital Forensics into Practice in Collecting Institutions”. White Paper. September 30, 2013.

[7] The Handle System, http://www.handle.net/index.html.

[8] InCommon, https://www.incommon.org.

[9] Helping Interdisciplinary Vocabulary Engineering, https://code.google.com/p/hive-mrc/.

[10] NCSA Polyglot, http://isda.ncsa.uiuc.edu/NARA/conversion.html.

[11] ClamAV, http://www.clamav.net/index.html

[12] Harvard Dataverse Network, https://thedata.harvard.edu/dvn/.

[13] Data Observation Network for Earth, https://www.dataone.org.

Beyond the Binary: Pre-Ingest Preservation of Metadata

Jessica Moran
National Library of New Zealand
P O Box 12349
Wellington 6001
+64 4 460 2862
jessica.moran@dia.govt.nz

Jay Gattuso
National Library of New Zealand
P O Box 12349
Wellington 6001
+64 4 474 3064
jay.gattuso@dia.govt.nz

ABSTRACT

This paper describes some of the challenges the National Library of New Zealand has faced in our efforts to maintain the authenticity of born digital collection items from first transfer to the Library through ingest into our digital preservation system. We assume that assuring the authenticity and integrity of digital objects means preserving the binary objects plus metadata about the objects. We discuss the efforts and challenges of the Library to preserve contextual metadata around the binary object, in particular filenames and file dates. We discuss these efforts from the two perspectives of the digital archivist and the digital preservation analyst, and how these two perspectives inform our current thinking.

General Terms

Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice.

Keywords

Digital archivists, born digital preservation, technical appraisal, ingest.

1. INTRODUCTION

The National Library of New Zealand (the Library) actively collects born digital heritage collections, including unpublished material such as manuscripts, personal papers, organizational archives, photographic archives, and oral histories. In 2008 the Library went live with the National Digital Heritage Archive (NDHA). The NDHA encompasses several ingest and access delivery components and utilizes the Ex Libris digital preservation software, Rosetta. The workflow for bringing these collections into the NDHA begins with an initial technical analysis by the digital archivists, then curatorial appraisal. Once appraised, collections are arranged and described, prepared for ingest, and ingested into the NDHA. This is a collaborative workflow with input from various stakeholder groups within the Library including digital archivists, digital preservation analysts, curators, and arrangement and description librarians. In the past few years our born digital collections have grown in both size and complexity. This growth has put stress on our workflows and challenged us to develop more rigorous pre-ingest and ingest processes. For example in the last year we have received four 1TB transfers, as well as two collections with 100,000 plus files. This is in addition to a number of hybrid collections which have included 50 or more physical carrier media items each.

At the Library digital archivists are responsible for the transfer of born digital manuscript and archives collections into the Library, as well as initial technical appraisal and preparation of collection items for ingest into the digital preservation system. The digital preservation analyst is responsible for technical assessment of digital content going into the digital preservation system, and troubleshooting digital content that fails validation checks. While both roles are informed by an understanding of both archival and technical considerations, the digital archivists serve as archival and content subject matter experts, while the digital preservation analyst is subject matter expert for technical concerns. The digital archivists and the digital preservation analyst work closely together developing ingest workflows. Bringing the two perspectives together allows us to design and develop robust workflows that better meet the needs of preserving the binary object, as well as contextual metadata around the objects such as filenames and dates. Working together gives us a better understanding of each other's perspective, a more holistic view of the digital preservation challenges, and ultimately allows us to have greater confidence in our ability to assert the provenance, authenticity, and trustworthiness of the digital content we are preserving.

This paper will outline, from the perspective of the digital archivists and digital preservation analyst, some of our challenges, particularly in the areas of filename and file dates. We will look at these two seemingly simple pieces of metadata, filename and file dates, and how we track and store that metadata from initial transfer to the Library until ingest into the NDHA from the perspective of the digital archivists and the digital preservation analyst. In the library and archival digital preservation environment we are familiar with the main types of metadata necessary for digital preservation such as descriptive, administrative (technical, rights, preservation) and structural. However we are also interested in file system metadata such as filenames and dates that are not embedded with the objects themselves, but rather are stored externally in the file system and are generated to track things like name, size, location, usage, etc.[1] Filename and date metadata would seem relatively basic metadata to capture and preserve, but in our experience, these two pieces of metadata have challenged us to think critically about what constitutes acceptable, reversible, and recordable change and where and how this metadata should be stored for preservation and later for delivery to users.

2. POLICY AND PLATFORMS

2.1 The Policy Context

The Library, in collaboration with Archives New Zealand (ANZ), has established a number of policies to support digital ingest and preservation activities. Of these, the Preconditioning Policy has had the greatest impact on the pre-ingest workflows. The policy states: “the diverse nature of digital content means that there are times when it is desirable to make changes to it before it is ingested into the preservation system.” These changes are classed under the term “preconditioning" and the policy describes the limits of change that can be introduced to digital content from the time it is under control of the Archives or Library to its being brought into the preservation system. These changes can include an alteration to (or addition of) file extensions and/or removal of unsupported characters in the file name. The undertaking of these preconditioning actions enables the Library to ingest more stable files into the preservation system’s storage database. Regardless of the change being made during this stage, the policy demands that one be able to demonstrate that actions will not affect the intellectual content of the file, all changes are reversible, and all changes must be documented with a system-based provenance note describing any changes made. [2]

2.2 The Business and Technological Context

The Rosetta digital preservation system and some of the business and operational rules around the Library’s use of the system influence our workflow decisions and requires some explanation. All files ingested into the preservation system go through the Validation Stack in Rosetta. The Validation Stack includes file-format identification (by DROID), file validation (by JHOVE), metadata extraction (NLNZ metadata extract tool and JHOVE, checksum generation (CRC32, SHA-1, MD5), and virus check (CLAM AV and F-PROT). This series of checks is run against all files in a Submission Information Package (SIP). During this validation process files within a SIP with missing or incorrect file extensions, some non-ASCII encoding of filename strings, and invalid dates will cause the SIP to fail validation and the entire SIP will be re-routed to the technical workbench area for assessment. Within this area of Rosetta there are limits on the actions that can be performed and the tools that can be used. The Library has found that when working with large unpublished collections that may require a series of fixes, it is more efficient to perform these actions outside the system prior to ingest. We also hope that ingesting files in this “stabilized” state will make them easier to report on and identify later. Further, we expect these preconditioning activities to make future operations and actions on the files, such as file format transformations for preservation, easier and less labor intensive. We are also conscious that any pre-ingest preconditioning activities we perform must be recorded systematically inside Rosetta.

For access and delivery of collection items, Rosetta acts as the link between a descriptive record maintained in a collection management catalogue system and the binary stream (digital object), or streams, that are being described. The Library’s collection policy, collection plans, and business rules inform the arrangement and description and access to unpublished digital content. While a discussion of these policies is outside the scope of this paper, it is worth noting that these plans affect our practice; essentially our preservation system and collection management system work together to deliver access to digital objects. This has led to the requirement to collect and deliver discrete digital objects that are linked with a descriptive record. To that end, our usual process is to cleave any digital objects from their host file system, because we are not typically interested in preserving the file system as a collection item. To maintain the collections as any number of file system sized items would break the cataloguing and delivery mechanisms we have developed and expect to continue using to deliver access to content in the future.

While we may create disk images as part of our pre-ingest technical appraisal and processing workflow, in most instances it will not be the disk images but the individual files we preserve in Rosetta. Therefore we have developed a workflow where we are able to individually address digital objects, marking items in or out of collection scope, marking individual items as in need of special attention, and ultimately addressing the whole collection as a collection of untethered digital objects, rather than as aggregate set of items whose individual needs cannot be easily addressed.

Accepting then that we have a requirement to transfer content from the host transport or storage media, we want to ensure that the record of the items held by the media as metadata is not lost during this separation process, and that we have captured that metadata in a format that can be easily used during appraisal, arrangement and description, and ingest. We also want this metadata included as part of SIPs and therefore retained within the preservation system. While there are any number of tools that address parts of this process it has been a challenge to create a workflow that could systematically and safely maintain this metadata in a way that is useful to our various stakeholders.

3. PRESEVING FILENAME METADATA

3.1 Filename information from a digital archivist’s perspective

The filenames of born digital objects, especially in manuscript and archives collections, provide us with information not about what the object is, but also what the creator might have been thinking during creation, and how the object relates contextually to other objects within a collection. For these reasons we want to be able to retain original filenames and deliver them back to researchers, even as we understand that within our preservation system Rosetta will assign new identifiers to files.

Born digital objects in manuscript and archives collections routinely come with notoriously non-standard naming conventions. Some of the issues that we encounter regularly include: older filenames with full stops in the filename, filenames with missing or incorrect file extensions, filenames with special and illegal or restricted characters in the filename, and filenames (and paths) that exceed the current Windows character limits. Other common naming issues we find include character encoding, soft hyphens, diacritics (especially in Māori language filenames), and other non-English language characters that neither our pre-ingest systems nor Rosetta can currently recognize correctly at ingest. If at all possible we will not touch filenames, however in these cases we will need to make a change to the filename in order to ingest the file into our digital preservation system.

In order to address these filename issues the Library adopted a preconditioning policy that sets the limits of acceptable change that can be introduced to digital content from the time it is brought into the control of the Library to the time it is ingested. For the digital archivists, who are responsible for the initial transfer of digital content into the Library, this gave us the framework for making acceptable changes to filenames where necessary. However, we have struggled with when, how, and where to document our changes and how to ensure that documentation makes its way into the digital preservation system. For example we create an original inventory of all files in a collection at our first contact with the material and then verify that that information is unchanged upon transfer to our pre-ingest storage location. These give us a snapshot of filenames and file path locations and establishes baseline fixity. We then identify any filename or extensions requiring preconditioning. [3]

Once we have identified filenames that need preconditioning prior to ingest, we must make those changes. In order to maintain the reliability and trustworthiness of our custodianship, we want to automate as much as possible not only the preconditioning actions we take, but also the recording of those changes, and their ingest into Rosetta as provenance metadata. For individual files or even small collections this is a relatively simple process. But once we began applying preconditioning to larger and more complex manuscript and archival collections, where for example we were processing thousands of files at a time, we discovered that we needed a better way to automate this process. [4]

3.2 Filename information from a preservation analyst's perspective

The preservation analyst provides technical assessment of digital content as it enters the digital preservation system and works especially with manuscript and archives collections when digital objects fail or are likely to fail technical validation checks. For the preservation analyst filenames pose an interesting and complex technical problem.

A filename can be reduced to a relatively arbitrary string that is used as a handle for an associated binary stream. They become a problem where an operating system (OS) or applications have reserved function for any of the characters or code points in the string that is being used. Different operating systems have different constraints for reserved characters, and over time operating systems have changed their rules.

For example in a Windows environment the characters: \ / : * ? “ < > |. are all reserved and have special associated functions meaning that you cannot name files with one or more of these characters. [5]

Linux systems have a different set of constraints restricting the use of forward slash and *NULL*, Mac OS prohibits the colon. [6] It's not just the OS that requires these restrictions; it is also the underlying file system that has some reservations in character usage.

These differences start to become an issue when the Library receives files from one OS and expects them to conform to the constraints of another. The main tools we have used while working with files prior to ingest are Windows based, and the preservation application has a couple of different file systems that inform the OS based filename character constraints. This means that what may be permitted in one environment, may not be permitted in another.

OS constraints are one part of the character issue we face, but we also need to deal with “non-standard” characters, like macronised vowels (e.g. ā, ē, ī, ō and ū) found in written Māori, one of the 3 official languages of New Zealand. [7] We cannot expect all our systems, users and processes to support the use of extended UTF-8 (or other such encodings) in filenames, and in fact, we often do not know what the original encoding was that was used to label the filename. We often encounter filenames recorded as “my□file.pdf” where the “□” glyph is essentially the OS recording software unable to represent the original character. We have also observed that different tools have different ways of addressing this problem of the presence of a code point it cannot decode. Some use question marks or a symbol like the above, others simply skip the offending character. But we must stabilize these filenames to ensure we are using correct, consistent, and valid filenames that can move safely between platforms.

File extensions are another component of the file name that is of interest to us. A file extension is often (incorrectly) used as a proxy for a meaningful file format identifier. Sometimes the format type and the file extension match, other times they do not. Operating systems have different methods of associating applications to various file types, one of which, particularly in the Windows domain, is the file extension. When the file extension is incorrect to a “normal” user it generally means that the file might not open with a suitable application, if at all. They may need to take some action to ensure that the file is properly associated with a suitable piece of render software, like changing the file extension manually.

Proper handling rules means that for digital preservation we need to treat files slightly more sensitively. We might want to know what the original file extension was as it is an important part of a file's provenance. Knowing that the creator labelled a binary stream “my_file.pdf” when it ought to be “my_file.tif” for example might be useful to a future researcher, so that when a the researcher is delivered the file as “my_file.tif” the information that it was originally named “my_file.pdf” is referenced elsewhere in the metadata for their information.

We also often encounter files missing extension altogether. In modern Windows environments this would not be common, but many manuscript and archives collections we work with were not created in a modern Windows environment. Further, some file systems and operating systems don't require file extensions and files often don't need an extension.

We made a policy choice to add or fix a file extension wherever possible, to prevent a file being delivered to a user without a valid file extension and to ensure that Rosetta has the right context clues so that the correct internal delivery mechanism is used.

This poses some challenging questions for us. What do we do when representing the original object to a user? Do we include the new file extension, or present the original string without the extension? How do we record any changes we've made and deliver that information to the user?

There is a further consideration. Rosetta has its own method of dealing with filename issues – in fact, at the first point of contact with a file Rosetta strips off the original filename, assigns the file a new filename, and retains the original filename in the Archival Information Package (AIP) metadata.

This acts as a useful normalizing process, because the files can subsequently be addressed by their new “clean” filename, but as soon as we want to deal with files in their original form, (to for example deliver to a user) we need to re-associate the binary stream with its original and potentially troublesome filename. The current method for recording these preconditioning provenance actions is to record a PREMIS provenance note that describes the state of the item before and after the intervention.

The provenance note is attached to the CREATION event, and is constructed using an in-house convention that was designed to support different types of provenance worthy interventions, including the changing of file dates, addition or change to file extensions, and the cleaning of special characters from file names. An example is offered below:-

Figure 1 - Example Provenance Note

```
<section id="event">
 <record>
  <key id="eventIdentifierType">Indigo</key>
  <key id="eventIdentifierValue">Indigo_1</key>
  <key id="eventType">CREATION</key>
    <key id="eventDescription">Provenance Note from
Indigo</key>
```

```
  <key id="eventDateTime">Wed Mar 18 12:47:24 NZDT
2015</key>
  <key id="eventOutcome1">SUCCESS</key>
<key id="eventOutcomeDetail1">002_File extension was
added: The file was submitted by the donor without a file
extension. On the recommendation of the Preservation
Analyst, a .wpd extension was added to this filename by the
Digital Archivist as it has been identified as WordPerfect for
Windows, version 5.1. </key>
</record>
```

This provenance metadata is expected to be used in the future to ensure we are able to provide both researchers and digital preservation staff with an accurate view of an object's metadata, including any preconditioning actions we may have performed on that metadata.

4. PRESERVING DATE METADATA

4.1 Dates from an archivist's perspective

Most born digital objects that come into the Library will have three dates: created date, last modified date, and last accessed date. However depending on what file system the objects were created on all of these dates may not be present. Ideally we would like to collect and preserve all three dates. It is important to note we would like to preserve these dates not so much because we believe that these timestamps gives us irrefutable information about when exactly the digital objects were created, modified, or accessed, but rather that they provide contextual information about how and when the objects were in use. They are, in essence, another piece of the puzzle used to confirm an object is what it says it is.

As the born digital collections being transferred to the Library increase in size and complexity, the amount of time from initial transfer to ingest into the digital preservation systems has also grown. This has the effect of increasing the time collections stay on pre-ingest storage, and may also increase the number of times the files are touched by digital archivists, curators, arrangement and description librarians, and digital preservation analysts. We have workflows and processes in place to mitigate risks associated with this, but we do not always have control over the underlying file system of our storage, and any changes that our information technology support might make. It is therefore important for us to understand and document both what and how the original timestamp dates were created, and how any subsequent movement of the files can effect these dates.

One of our first steps during a transfer is to view the transfer media using a write blocker and use tools to capture an inventory of the digital object's basic file system metadata including file name, file location, size, and dates available from the storage media on which the collection was transferred to us. We then must transfer the files to a pre-ingest location without affecting these dates. We found we were having to cobble together a number of tools, and therefore steps during the transfer process. We also understood that every time we wrote a file to a new file system these timestamps were subject to change, and indeed, every time we interact with a file, we run the risk of altering the timestamps.

We needed to adopt processes to ensure we did not inadvertently affect the file metadata and can ensure the authenticity and reliability of the digital objects under our control, while at the same time working in an environment that allows for multiple people to work with objects, and prepare the files for ingest into the digital preservation system.

One solution is the creation of forensic disk images as a first step in the transfer process, to essentially wrap all the digital objects in an image file that records file system and related metadata for later use. [8] At the Library we have adopted many lessons in workflow and digital handling from the world of digital forensics. [9] However, as discussed above our use case will in most instances call for extracting and loading the individual files to the digital preservation system. In these cases, even if we do have a disk image and all the relevant file system and individual file metadata, we again find ourselves in the same dilemma, that is: how do we preserve the original timestamps, even if the files themselves have been moved? And if the timestamps associated with the original digital objects are not the timestamps associated with the object at the point of transfer to the Library, how do we ensure that the original information travels with the object? While this information can be captured in a DFXML file, once we extract individual files from the disk image, that metadata becomes disassociated from the individual files and requires more steps to remarry inside the digital preservation system.

The problem of dates become further complicated when working with older born digital collections that come into the Library in less than ideal condition. In a recent example we processed a manuscript collection that came into the Library on 3.5" high density floppy disks. On our first inventory of these files we noted that the dates associated with some files appeared to be broken. That is, on about a third of the files, rather than the expected 1990-2000 era created and modified dates, we instead saw dates that ranged from 2032 to 2066. These errant dates presented a number of questions for us. First, did we want to preserve these dates as they appeared? Certainly this was how they arrived in the Library and preserving the dates, even if they were demonstratively wrong, seemed a legitimate strategy. However the initial attempts to ingest the files with these dates failed as our digital preservation system rejected the files with dates from the future as invalid.

We then investigated these files in more detail. Returning to the disk images, we noted that the erroneous dates were present on the disk images. We investigated whether it was an issue with the original file system and how the files were written to disk and then being read by our OS. However, the erroneous dates were present regardless of what OS we used.

We next asked ourselves, if we could not discover the original or "true" timestamps, how we might change those dates in a trustworthy manner, maintaining authenticity. In order to not violate our preconditioning policy, we needed to ensure any change we made did not change the intellectual content of the object, was reversible, and that we provided sufficient documentation of what changes we made to the objects and why. Finally, if we were going to have to change the timestamps, what would be the most appropriate date to change the files to? Eventually we concluded that if we were going to make changes to timestamps we should make them as transparent as possible to future users. Thus we chose to change all erroneous dates to a current 2015 date, in the hope that such dates would alert future users to the discrepancy in the purported dates in both the content and descriptive records and the file date stamps, and hopefully lead them to further investigate the provenance notes in the preservation metadata.

4.2 Dates from a digital preservation analyst's perspective

The collection described above was a particularly troublesome set of files, and worth describing in more detail.

Before we look at the collection, it's worth restating some of the principles we have adopted when handling files.

- We expect a file to arrive at the Library with accurate dates. This includes created, last accessed and last modified dates. Some files / file systems include other dates, such as last printed.
- We understand that not all OSs support all those dates, and not all files have this metadata available. For example earlier Linux versions specifically did not support the notion of a created date as it is not required by POSIX the standard (the basis of an underlying file system) [10]
- We expect to be able to handle files without changing those dates – we consider them to be a valuable part of the item.
- We understand there is a particular issue around the "created" date of a file in the archival context. (Should this date reflect the date a clone of a bitstream was made, or the original bitstream?)
- We understand that different types of date / time have different resolution, ranging from milliseconds to whole days. [11]
- We understand that moving a file changes the OS metadata differently to copying a file.
- We have a practice of only working on copies of original files wherever possible, especially when testing procedures.
- We have a practice of touching the original file as little as possible and only as much as needed to get the file into the preservation environment
- We find that at present there is no perfect tool or process that we can use that ensure we can operate with absolute comfort.

These rules are not all set in stone, and we have been working on our business practices and tools to help us align our practice with our policy as much as possible. Each time we encounter a problematic collection, our model shifts slightly to accommodate new learnings or specific nuances of a file or set of files.

In the example discussed above, we knew we had a problem ingesting these files when one of our automated tools returned an error.

In this case we were using the Python ZIP library [12] to coerce the collection into a single zip file, ready for automated ingest into Rosetta (via a process we call the csv ingest method).[1] The scripts we had written returned an error because it could not handle file objects created after the current date.

This is not really surprising as conceptually the files should not exist if the file creation date is assumed to be true, however there are no technical controls over the setting of such date, and various methods can be used to change any dates associated with a binary stream in a file store. [13]

In our example set, we found that a third of the dates we could find were incorrect, and about half the number of dates we thought we ought to have (1 x created date, 1 x last modified and 1 x last accessed date per file, for 1,101 files) was missing from the source set.

We noted that the missing and incorrect dates were scattered amongst the collection, which was originally housed on seven 3.5" high density floppy FAT12 formatted disks.

The incorrect dates fell into two groups, a set that ranged from 2032 to 2035 and a set that ranged from 2062 to 2066. The original set ranged from 1999 to 2003.

It was notable that the original dates spanned a four-year period, as did the wrong dates, just with a relatively consistent offset.

Our working assumption for some time was that if we could find the offset, we could assert the original date with some confidence. Being copies of emails and travel schedules many of the affected files actually had dates as part of the intellectual content. This meant we could tell if we were in roughly the right time period.

That approach held for a while, at which point a discovery that the time portion of the date was equally affected.

This left us in some doubt that the original date could be recovered, and in conclusion we returned to the idea that we would set the dates to something obviously (to us) wrong and ingest the files with provenance information that documents the decision and justification for our actions.

This problem is not dissimilar to those we face with filenames, and it follows that we might want to use a similar approach. If we capture the filename and dates that are available to us at the first point of contact with a file, and store them in a suitably convenient way, we can act with more confidence in changing this type of metadata.

5. DEVELOPING BETTER TOOLS

5.1 One possible solution

One general characteristic of digital curation work is finding oneself in a situation where there is both a preponderance of tools and no one tool or even suite of tools, that meets all our needs. In the Library's case what we needed was a tool to help us automate the original and any subsequent transfers of born digital content, ensure the capture of original filename and date metadata and any preconditioning actions we performed, and at the same time create a log of that activity that is auditable and both human and machine readable.

5.2 Our requirements

Working together, the digital archivists and digital preservation analyst developed the Library's requirements. These requirements included needing a way to create an inventory of each file on a piece of media. This inventory should include the original filename, file path location, an MD5 hash of the binary object, the file date created, date last modified, date last accessed, and the file size. But we did not just want that information, we wanted the ability to take all that information, move all the files to a new storage location, recheck everything and confirm that all the data points were the same and that no changes were made in the course of the transfer, or if changes were made, to record those changes. Finally, we

[1] CSV ingest is a method of ingest in which we create a ZIP of all the files to be ingested, as well as a csv manifest of all the SIP metadata to support the ingest directly to Rosetta. Rosetta then builds the SIP and ingests the files automatically.

needed a way to compare these inventory logs if a set was moved more than once, as is required from time-to-time.

Based on previous experiences with other file copying tools that either failed or silently made changes to filenames with reserved, and non-standard characters, and our prior transfer experience, we knew we would continue to regularly receive files with these issues. Therefore we also needed a way to automatically remove any restricted characters and record those changes at the time of the initial transfer.

Essentially what we wanted was a tool that can meaningfully capture this metadata, keep that metadata linked to the original objects, and be integrated into our existing workflows and systems. By bringing our archival and technical perspectives together, we have begun working on the early stages of a tool that can do this for us in a way that meets both our needs and the needs of our systems. [14]

The underpinning requirements that we are attempting to address cover the following:

- Must be easy for relatively non-technical users to deploy
- Should result in the transfer of all files found in a mounted file store, starting at a nominated folder (recused through the extents of all child folders to a nominated location elsewhere
- File system structure should be mirrored
- File fixity is recorded and maintained in the copied item
- File dates are recorded and maintained in the copied item
- Filename and paths are allowed to be sanitized as long as done in accordance with preconditioning policy.
- Boolean comparisons of source and destination are recorded for all data elements (file names, paths, fixity and dates)
- Elements are recorded in a way that allows accurate metadata and provenance information to be captured with binary objects and used in creation of ingest SIPs.

Most of the above is present in the script we developed, and we have used an early version of this script to confidently move several terabytes of data, and further to assist in the automatic ingest of complex collections into Rosetta.

We have encountered a number of interesting occasions where it was not possible to reassert some of the original metadata on the copied file. The script uses a metadata preserving method of copying a binary from A to B. It collects accessed date, modified date, and created date as applicable and records these in a log for that item. Further, it checks that the dates found on the copied item are the same as the original item. If they are different, it attempts to assert the original dates onto the copied file. It finally checks the two items again, and if they are still different this discrepancy is recorded as another data element. In the case of dates failing to be maintained for the copied item, the original dates are captured before we try and copy the item, and the Boolean flag is set to clearly indicate we need to create a provenance event for changes in dates to ensure an integrious record is maintained.

The steps enumerated above are an improvement on our existing practice, and result in data that can be used to confidently move content from location A to location B, capturing accurate metadata from any operating system / file system that it touches along the way, and specifically helping to drive an increasingly automated pre-ingest workflow where applicable.

Our ongoing questions concern the extent to which delivery of objects from the digital preservation system should include not just a stated assurance that this is an authentic binary object, but also proof of that integrity and authenticity through delivery of the associated metadata. Attempting to answer these and other questions that are sure to arise will require us to continuing working together and learning from each other's perspectives.

6. ACKNOWLEDGMENTS

Our thanks to Leigh Rosin (Digital Archivist, National Library of New Zealand) who was part of the many conversations that informed this paper.

7. REFERENCES

[1] Rogers, C. 2015. Diplomatics of born digital documents – considering documentary form in a digital environment. *Records Management Journal* 25, 1 (2015), 6-20. DOI= http://dx.doi.org/10.1108/RMJ-03-2014-0016

[2] Joint Operations Group, Policy (JOGP), Department of Internal Affairs, 2012, "Digital content preconditioning policy," pp.1-2. http://ndha-wiki.natlib.govt.nz/assets/NDHA/About-Us/Strategic-Partnerships/Digital-Preservation-Policy-Manual.pdf .

[3] Rosin, L. 2014. Applying theoretical archival principles to actual born-digital collections. *Archive Journal 4* (Spring 2014). http://www.archivejournal.net/issue/4/notes-queries/applying-theoretical-archival-principles-and-policies-to-actual-born-digital-collections/ [accessed 1 July 2015].

[4] Rosin, L. and Smith, K. 2014. Then and now: the evolution of digital preservation and collecting requirements over a Decade. In *Proceedings of the 11th International Conference on Digital Preservation* (Melbourne: State Library of Victoria, 6-10 October, 2014) https://www.nla.gov.au/sites/default/files/ipres2014-proceedings-version_1.pdf .

[5] Naming Files, Paths, and Namespaces, https://msdn.microsoft.com/en-nz/library/windows/desktop/aa365247%28v=vs.85%29.aspx, [accessed 17 April 2015]

[6] File Naming Conventions in Linux, http://www.linfo.org/file_name.html [accessed 17 April 2015].

[7] Korero Maori, Tohuto – Macrons, http://www.korero.maori.nz/resources/macrons.html [accessed 17 April 2015].

[8] Lee, C.A., Woods, W., Kirschenbaum, M., and Chassanoff, A. 2013. *From Bitstreams to Heritage: Putting Digital Forensics into Practice in Collecting Institutions*. White Paper. September 30, 2013.

[9] Kirschenbaum, M.G, Ovenden, R. and Redwin, G. 2010. *Digital Forensics and Born-Digital Content in Cultural Heritage Collections.* Council on Library and Information Resources, Washington, D.C. http://www.clir.org/pubs/reports/reports/pub149/pub149.pdf

[10] Unix &Linux, How to find creation date of file? http://unix.stackexchange.com/questions/91197/how-to-find-creation-date-of-file [accessed 17 April 2015].

[11] Corbet, J. (2007) "Once upon a time." http://lwn.net/Articles/244829/ [accessed 17 April 2015].

[12] Python Standard Library, 12.4 Work with ZIP archives. https://docs.python.org/2/library/zipfile.html [accessed 17 April 2015].

[13] Python – change file creation date, http://stackoverflow.com/questions/887557/python-change-file-creation-date [accessed 17 April 2015] and How do I change the file creation date of a Windows file from Python, http://stackoverflow.com/questions/4996405/how-do-i-change-the-file-creation-date-of-a-windows-file-from-python [accessed 17 April 2015].

[14] Safe_mover.py, https://github.com/jayGattusoNLNZ/Safe_mover [accessed 17 April 2015].

Characterization of CD-ROMs for Emulation-Based Access

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

Tobias Steinke
German National Library
Adickesallee 1
60322 Frankfurt, Germany
t.steinke@dnb.de

ABSTRACT

Memory institutions have already collected a substantial amount of digital objects, predominantly CD-ROMs. Some of them are already inaccessible with current systems, most of them will be soon. Emulation offers a viable strategy for long-term access to these publications. However, these collections are huge and the objects are missing technical metadata to setup a suitable emulated environment. In this paper we propose a pragmatic approach to technical metadata which we use to implement a characterization tool to suggest a suitable emulated rendering environment.

General Terms

Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows.

Keywords

Emulation Characterization Tools

1. INTRODUCTION

Memory institution have already accumulated a substantial amount of digital artifacts. For instance, the German National Library (DNB) has been collecting German publications on various data carriers since commercial publication began in the end of the seventies. There are all kinds of data carriers for many different computer systems in the magazines of the DNB. The estimation of the number of stored digital carriers is about 500000. It is not exact as the cataloging of these publications was not consistent in the early years.

Currently, a user in the DNB's reading room can order a data disk via catalogue, which will then be prepared for usage in a virtual drive. These CD-ROMs, however, need to run in the DNB's current desktop computer environment (currently Windows 7). In some cases the CD-ROMs still work but other CD-ROMs fail or may fail soon. Hence, the current access workflow lacks a future proof strategy. Furthermore, similar collections of digital publications with a dynamic character, for instance web archives, also need a future-proof access solution.

Emulation has been shown to be a useful and flexible approach to access legacy software collections [2, 15, 4]. However, in libraries and their classical reading-room scenarios, the development and maintenance of tailored emulation solutions is difficult. Especially for very large collections, analyzing and evaluating each object individually is not a feasible solution [13]. Based on these findings, the EMiL [1] project aims to develop an environment for library reading rooms that addresses the challenges of complex objects with a flexible emulation framework. The primary focus for the DNB is on providing access to multimedia CD-ROMs of the 1990s and 2000s, but the EMiL system might be used for other collections later on as well. Automation is essential, since the huge number of objects prevent manual handling of each object, in particular in view of the determination of technical metadata required for emulation.

The default scenario for the DNB is rather simple: Users in the DNB's reading room may use specific workstations for research purposes. If any of the catalogue's entry is a multimedia publication originally published on a data carrier, a click on a link initiates the transfer of the data carrier's digital image from the digital preservation system to the EMiL system. Then, the user gets access to the content of the publication using a suitable emulation environment. No additional interactions or decisions by the user should be required during this process. Hence, for this rather simple workflow it is crucial that the EMiL system is able to automatically determine the "best" available emulation environment.

Therefore the development of a trustworthy characterization tool and a flexible management of emulation environments are key tasks in the project EMiL. In this paper, we present design, implementation and evaluation of a CD-ROM characterization tool as well as necessary technical metadata and technical components.

2. PROBLEM DEFINITION

As outlined in the previous section, EMiL aims at re-enacting a multitude of digital objects using emulation technology.

[1] Emulation of Multimedia-objects in Libraries, http://www.multimedia-emulation.de/

However, these objects cannot run without suitable technical environments. These environments typically include a (virtual) computer hardware, an operating system and sometimes additional applications, all of which form the complete emulation or rendering environment for a digital object. Due to the many ways hardware and software can be combined, not any such environment can be used to render a specific object. Instead, the EMiL software framework has to find environments suitable for a specific object based on the technical aspects of this object.

When objects like CD-ROMs have been acquired and ingested into a memory institution's catalog, typically descriptive metadata has been gathered and associated with the object. Even though these metadata entries may also describe technical aspects of the object, (re-)using this data to identify and prepare a suitable rendering environment is difficult. Listing 1 shows an example of technical system requirements of a CD-ROM published in the late 90s. These technical descriptions were originally designed to guide potential buyers, but from today's viewpoint most of the information is irrelevant (computing power and memory and disk sizes have risen by magnitudes). A much bigger problem is that there was no standard or schema in describing system requirements; hence, using this information as technical metadata in an automated way is difficult.

Listing 1: System requirements as posted on the CD-ROM box or booklet.

```
Windows 3.1x, 95, NT 4.0
IBM Compatible PC
80486 Processor or higher
Minimum: 8 MB Memory (RAM)
10 MB free hard disc space
Minimum DIN/ISO 9660 CD-ROM drive
MSCDEX 2.21
Mouse
```

The KEEP Emulation Framework [11] provides file characterization in order to determine a ViewPath [14]. Based on a file format, a rendering application is chosen; using the application, an operating system and finally an emulator is determined [1]. The information required is stored in the framework's database. This file-based approach can be used very efficiently to cope with certain types of digital objects (digital pictures, text documents, etc.). File types can be automatically characterized which allows to select the corresponding viewer software.

This single-file characterization approach falls short on more complex objects. Object types that consist of a collection of different files and formats that are wrapped into a single container format (e.g. ZIP or ISO9660 images) cannot be classified with a single file format. Today's file characterization tools may recognize the container format, which is, however, of limited use for a useable access strategy. Even if the container's content is analyzed file by file, this approach can only provide a file-by-file re-enactment and loses the collection-type character of the object. This is true especially for interactive CD-ROM publications (e.g. multimedia productions, interactive educational content, encyclopedias etc.) where just providing access to individual image and text files is not sufficient if the original digital object provides a rich application that guides the user through the available material.

Even if an object's "viewer" software or its ViewPath is known, this information is neither unambiguous nor sufficient. Usually, executing a specific application is not possible without prior installation, neither is creating such an environment in an automated way. Similarly, an emulator system required to run such software needs to be set up and configured properly. Furthermore, there exists an (almost) unlimited number of potential software environments satisfying an object's technical requirements. Hence, in a practical scenario an image archive with associated metadata is required which provides means to search for a matching environment.

In earlier work we have presented workflows for manual ingest of CD-ROMs, i.e. the user individually chooses from a list of available rendering environments [9]. If no suitable environment is available, a new environment has to be built, for instance, by installing additional software. The result of the CD-ROM ingest workflow is a semantic link between a digital object and the technical description of an emulated system environment able to render a certain object. This approach is viable and useful for instance in the digital art domain since the rendering quality can be evaluated. However, good knowledge of each CD-ROM is required to be able to choose or create an appropriate rendering environment. Traditional memory institutions, however, hold large collections of different digital objects. A manual (re-)ingest of these objects is labor-intensive, as this workflow requires an in-depth analysis of each object to determine its rendering environment. For a majority of these objects (e.g. for supplementary CDs) the total costs would probably not be justifiable.

Our goal is to support and automate this process as much as possible. From a set of available rendering environments, only a few are suitable to render a given object. This matching process requires a characterization tool as a prerequisite. Furthermore, with respect to the memory institution's large collections, gathering information on required environments covering (most of) the collection's objects is necessary to build a comprehensive image archive to be chosen from.

3. TECHNICAL METADATA

The problem definition and requirements show that technical metadata describing the capabilities of the rendering environment is necessary. Several models for describing a computing environment have already been developed. The PREMIS data dictionary [12] provides a semantic *environment* entity to describe rendering environments, which has been recently reworked and extended to improve expressiveness, in particular in emulation use-cases [7, 6]. Similarly, The Trustworthy Online Technical Environment Database – TOTEM [8] provides a comprehensive data model to describe environments in great detail. Furthermore, tools have been proposed to determine environment information, for instance capturing a digital object's runtime dependencies [5].

The major trade-off, from a practical perspective, is choosing between the level of detail and the ability to re-use environments with new emulators, different objects or usage contexts. A small number of generic environments to render a large number of digital objects is preferable, as it reduces the burden of preservation planning. With detailed and very specific environment descriptions, re-use of environments becomes less likely. Also the complexity of associated tools matters. For instance, with more details, it becomes more difficult to design a matching algorithm that links rather generic emulator software with specific hardware requirements.

For this reason, we pursue a constructive approach, with less focus on formal modeling yet. The primary purpose of our technical metadata is to describe environments to render digital artifacts but also to describe environments such that preservation planning for emulation-based preservation strategies become possible. In order to integrate emulation related metadata seamlessly into a memory institution's existing preservation systems, a formal model and possibly encoding in a standard metadata language (e.g. PREMIS) may be required. This, however is left for future work.

A complete set of metadata required to re-enact a virtual environment is called an emulation environment and can roughly be divided into two parts, a hardware environment and a software environment. The former describes the technical features of an (emulated) computer system, while the latter describes the software utilizing the computer hardware, usually in form of a virtual disk image. A digital artifact typically poses (abstract) requirements on its rendering environment, e.g. a Windows operating system (Windows 95 or newer) with a set of applications installed and sound support enabled. However, specific requirements on the sound hardware are rare. Hence, an artifact is (tightly) linked to one or more software environments, but in general indifferent to the underlying hardware environment (as long as appropriate functionality is provided). The hardware environment, on the other hand, is selected based on the aggregated software environment's requirements and hardware's capabilities. To use hardware features, usually driver software needs to be installed and/or the operating system requires configuration. A software environment thus limits the choice of useable hardware configurations.

The separation of hardware and software environment descriptions allows to change the underlying (emulated) hardware without reevaluating (a huge number of) digital artifacts. This, makes it easier and more cost effective to cope with technological changes. Operating systems (in particular old ones) do not change their hardware requirements or their features over time. Hence, a particular software environment description can be considered as constant (if complete), especially regarding the hardware interfaces used. Even though the link between a specific software and hardware configuration is also stable in the short run, emulators are also prone to a software life-cycle and will be technically obsolete at some point. Then, if new emulation software is required, a new hardware environment description is created, describing individual emulated hardware components. If the connecting interfaces between hardware (hardware component description) and software (driver and operating system configuration) are made explicit in the technical metadata, all affected software environments can be determined, and the search for new emulators can be guided. If no perfect match is found, the necessary adaptions of affected software environments can be predicted in an automated way. This allows to focus preservation planning activities on monitoring the links between software and hardware environments.

Listing 2: General structure of a software environment description.

```
<swEnvironment>
 <id>..</id>
 <description>...</description>
 [...]

 <binding>
  <url>nbd://my.host?exportname=disk.img</url>
  <md5sum>...</md5sum>
  [...]
 </binding>

 <systemConfiguration />

 <softwareCollection />

</swEnvironment>
```

3.1 Software Environment

A software environment description's primary purpose is to describe a computer system's software setup. In the case of emulation workflows, these setups can be found in the form of disk images, representing a virtual hard disk to be used with an emulator. Listing 2 illustrates an `swEnvironment`'s general structure and main elements.

The first important element is a data `binding` definition, referring to the location of the disk image. Data bindings define volumes that can later be used to emulate a medium but only represent the bare data or data source. This element may contain further information, such as fixity information, format and file-system of the container.

The second main element of a software environment describes its relation to a potential hardware environment description. The `systemConfiguration` tag describes if there is an operating system installed and configuration of (required) hardware dependencies. The `osConfiguration` element describes the operating system as a `swComponent`, in particular with its rendering capabilities. For instance, a Windows 98 installation is able to render (execute) Win32 executables (`x-fmt/411`) (native format) but may also run Win16 (`x-fmt/410`) and MS-DOS executables (`x-fmt/409`) (import formats). In our implementation, we currently use PRONOM IDs (PUID) [3] to specify the supported rendering capabilities, if available. While the format specifications can be in any form, we chose PUIDs both because of its simple scheme and to be able to (re-)use its file format and software descriptions. While the PRONOM registry is far from complete, in particular with respect to software descriptions, it provides a viable (and well known) starting point, which could be quickly extended with a growing emulation community.

In order to describe a software's rendering capabilities, we chose a slightly different structure. While PRONOM uses *create* and *render* categories to associate file formats to a software description, we distinguish between *import*, *export* and *native* formats. For our use-cases it is helpful to be able to choose between applications rendering their native format and applications which are able to render a format to a certain extent. Software usually has a native data format that can be rendered without losing information. At the same time, it often allows some sort of interoperability with other software and thus provides at least partial support for other formats. For instance, OpenOffice.org natively supports OpenDocument but is also able to open Word documents, even if the rendering quality may be imperfect. This concept also provides a path to migrate independent data objects to other formats, e.g. to migrate a Word Perfect file to current Office Open XML.

As a specific software environment, in particular the embedded operating system, does not run on any hardware, the system configuration section also contains information about specific hardware configuration. An operating system's configuration and its (additionally installed) drivers define a set of ”expectations” on the underlying hardware system. This `hwConfiguration` description can then be matched against hardware environment descriptions. Listing 3 illustrates a simple system configuration of a typical Windows 98 installation.

Listing 3: A software environment's system configuration description.

```
<systemConfiguration>
 <osConfiguration>
  <swComponent id="x-sfw/37">
   <description>Windows 98 SE</description>
   <nativeFormats>
    <fmt puid="x-fmt/411" />
   </nativeFormats>
   <importFormats>
    <fmt puid="x-fmt/410" />
    <fmt puid="x-fmt/409" />
   </importFormats>
  </swComponent>
 </osConfiguration>

 <hwConfiguration>
  <hwcomponent class="audio" device="sb16">
   <param key="irq" value="5"</param>
   <swComponent id="x-driver/99 />
  </hwcomponent>

  <hwcomponent class="storage" device="piix3"
      id="ide_1" />
  <hwcomponent class="storage" device="cdrom">
   <param key="controller" value="ide_1"</
       param>
 </hwConfiguration>
</systemConfiguration>
```

The metadata associated with a `softwareCollection` extends its rendering capabilities, i.e. the type of file that can be used within this software environment. Software metadata could be kept directly in the software environment description but it is preferable to use only references to the associated software archive or technical registry. This way, software properties, in particular, license information, where conditions may change over time, can be centrally maintained and evaluated on-the-fly.

Listing 4: Excerpt of a software collection description.

```
<softwareCollection>
 <swComponent puid="x-sfw/68">
  <description>
   WordPerfect Office V.11
  </description>
  <nativeFormats>
   <fmt puid="x-fmt/44" />
   [...]
  </nativeFormats>
  <importFormats>
   <fmt puid="fmt/125" />
   [...]
  </importFormats>
  <exportFormats>
   <fmt puid="fmt/97" />
   [...]
  </exportFormats>
 </swComponent>
 [...]
</softwareCollection>
```

3.2 Hardware Environment

The basis for running any software environment is either a physical or an emulated computer system. The hardware environment description is quite similar to the software environment, as it defines a set of available hardware interfaces that make up a computer system.

A hardware environment is typically defined through its basic architecture (e.g. x86 compatible PC) and its specific hardware features. A list of list of `hwComponents` describes available hardware components, which a software environment is able to utilize. Each `hwComponent` represents specific hardware and describes configuration options, if necessary. For instance, Listing 5 lists the audio cards supported by the QEMU i386 system emulator. The system environment's `hwConfiguration` from Listing 3 then chooses one of these cards (Sound Blaster 16) and configures it accordingly.

The source of this environment description is usually the emulator's manual, describing the emulator's capabilities. By making the list of supported hardware explicit, different emulators can be compared. For instance, Virtual Box supports only three sound cards (Sound Blaster 16, AC97 and Intel HDA), VMWare supports only Sound Blaster and Intel HDA cards. The same applies to other hardware components like network and graphics cards and storage controller. Based on this information, one can produce guidelines to guide the development of software environments, in particular operating system installation and its configuration. Hence, the software environment from Listing 3 should be compatible with QEMU, VirtualBox and VMWare with respect to audio hardware requirements.

Every hardware environment available in the EMiL framework is backed by an emulation component implementation (cf. [10]). Each component is able to parse a `systemConfiguration` element and translate its requirements into native configuration for a specific emulator software. Table 1 shows an overview of EMiL's supported emulators.

Listing 5: System description excerpt as featured by a current QEMU emulator.

```
<hwEnvironment>
 <id>...</id>
 <name>QEMU i386</name>
 <architecture>x86</architecture>
 [...]

 <hwComponents>
  <hwcomponent class="audio" device="sb16">
   <param key="irq" value="5"</param>
   <param key="irq" value="7"</param>
   [...]
  </hwComponent>
  <hwComponent class="audio" device="ac97"/>
  <hwComponent class="audio" device="es1370"/>
  <hwComponent class="audio" device="hd"/>
  [...]

  <hwComponent class="storage" device="piix3"
      id="ide_1" />
  <hwComponent class="storage" device="cdrom">
   <param key="controller" value="ide_1"</param
       >
  </hwComponent>
 </hwComponents>
</hwEnvironment>
```

4. DESIGN & IMPLEMENTATION OF A CD-ROM CHARACTERIZATION TOOL

Based on the DNB's reading room scenario, the access workflow starts by requesting an object from the library's catalog. The goal is now to determine a suitable environment for this object automatically. To this end, a software environment suitable for rendering the object at hand has to be determined. This matching process has to be based on the requirements of the object and its expectations about the environment it runs in, e.g. a certain operating system version or support for the file type of the digital object.

Table 1: List of EMiL standard environments

Operating System	Arch	Emulator	Alt. Emulators
MS-DOS	x86	QEMU	Dosbox, VBox(VX)
MS Windows 3.11	x86	QEMU	Dosbox, VBox(VX)
MS Windows 9x	x86	QEMU	VBox(VX)
MS Windows XP	x86	QEMU	VBox(VX)
Linux i386	x86	QEMU	VBox(VX)
Apple II	MOS Tec	PCE	vmac-mini, MESS
Apple System 7	m68k	BasiliskII	MESS
Apple System 8	ppc	Sheepshaver	MESS
Apple System 9	ppc	Sheepshaver	
Amiga	m68k	x-uae	MESS
C64	MOS Tec	VICE	MESS
Atari	m68k	hatari	MESS

4.1 Building an Image Archive

The test collection for the EMiL project consisted mainly of interactive, runnable software objects rather than bare document formats (like images or Word documents). CD-ROMs were usually made for the mass market and are therefore mostly self-contained, i.e. if additional software was required, which is not already part of the CD-ROM image (Quicktime or Acrobat Reader are popular examples). Hence, the most important feature for a software environment is to run applications made for a specific operating system and computer architecture.

In order to provide suitable runtime environments, firstly, a list of necessary standard environments has to be compiled. These so-called standard environments provide a basic operating system installation and configuration, so that there is least one hardware environment (i.e. emulator) satisfying the resulting software environment's `systemConfiguration`. We have chosen executable file formats as our primary identifier for the gathering runtime requirements.

Executable file formats, however, are usually not very precise, because they are designed as a very basic interaction point between the operating system and CPU code. For instance, the portable executable (PE) file format used on Windows operating systems has been stable since Windows NT 3.1 and Windows 95 until the recent introduction of the 64-bit architecture and is still today used for non-64-bit binaries on Windows. However, a PE binary designed for Windows 95, is not guaranteed to run on Windows 8 and vice versa. To cope with this dilemma, we implemented a second classifier to the matching mechanism. Depending on the file's timestamp we can distinguish between different epochs of an operating system or software package and thus ”authentic” software environments.

Using these two classifiers, the binary file format and the epoch, we analyzed all images in the sample collection and produced a set of operating system and architecture (cf. Table 1). Column 3 and 4 of the table also show emulators that provide the technical features to emulate these environments. This classification step showed that the object collection provided by the DNB can be re-enacted using comparatively few environments, in particular if compared to the large number of objects.

Once the basic set of software environments required to access the digital objects has been determined, these software environments have to be created. This process includes the installation of an operating system, configuring it accordingly and installing required drivers for e.g. network or sound support. As a hardware basis, a typical configuration of the corresponding epochs computer systems is used, respectively. The result is a freshly made software environment with its characteristics known and described with a complete set of technical metadata. These software environments, including the disk images of the installed operating system, are then ingested into the project's image archive. Environments can then be used to be matched against the analysis results of individual objects. As they were created according to the requirements gained from the previous analysis of the whole collection, we can ensure that all objects find a matching software environment or we can determine precisely which objects lack specific features not (yet) included in the image archive.

4.2 Managing Software

For some objects in the sample collection, the process described previously failed due to our focus on executable binary formats. For instance, one object contained only a single PPT file with all other media directly embedded. This object, obviously, contains no executable binary format and

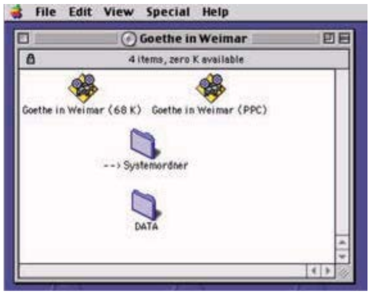


Figure 1: Content of a hybrid CD-ROM opened on an Apple Macintosh computer system.

thus does not require a specific operating system per se. However, the PPT file requires another piece of software that is shipped neither with the operating system nor the digital object, namely Microsoft PowerPoint.

To add to the existing base environments and provide auxiliary software environments that provide support for further file formats, it is possible to create derived environments within the image archive. To facilitate this task, a separate software archive can be provided that contains installation media for several additional software packages, e.g. MS Office. Accompanying these installation media, there is also metadata describing the additional native, import and export file formats provided by the software. The installation process, then, is similar to the re-enactment of a digital object: a media container and file system analysis yields the required software environment that this software can be installed on. After the installation process, the modified image can be ingested back into the image archive with the updated metadata as a derived environment based on the original base environment.

Internally, only the modified data blocks of the hard disk image are stored while the original blocks just reference the original base image. Therefore, the derivative remains dependent on the availability of the original environment. Due to this link, it is also possible to precisely tell which software environments may require looking into, once a base environment changes for some reason (e.g. because new emulators require new drivers to be installed).

Using these newly created derived software environments, the object only containing a PPT file can be started by searching for the respective PRONOM IDs. Now, the derivative with MS Office will be a matching result and the object can then be rendered using PowerPoint. Having such software available individually also provides means for ensuring the license limits of certain software is not exceeded. For instance, a memory institution may have a large number of operating system licenses but only a very limited number of special-purpose software that not everyone needs. Rendering all objects that do not require that special-purpose software using software environments that do not contain them allows for a larger number of parallel users.

4.3 On-the-fly Object Characterization

Once the image archive contains software environments and corresponding metadata, we are able to match the requirements of a digital object requested by the access workflow against the software environments and find a suitable rendering environment. As the original text-based requirements cannot be used to automatically and reliably find a rendering environment, we have to provide a different matching process that relies on the object's actual contents. This matching process is outlined in Fig. 3.

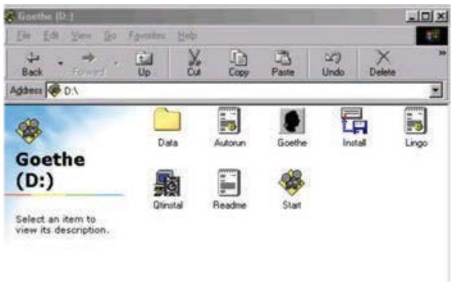


Figure 2: Content of the same CD-ROM opened on a Windows computer system.

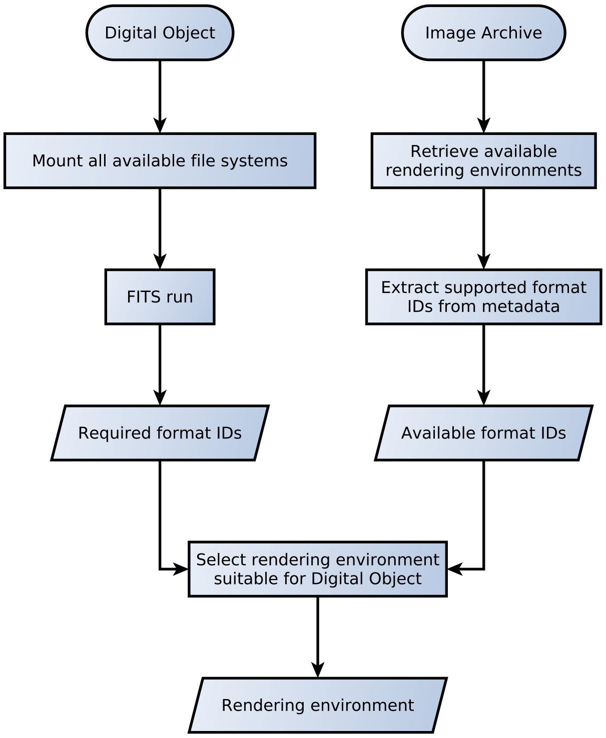

Digital Object
Image Archive
Mount all available file systems
Retrieve available rendering environments
FITS run
Extract supported format IDs from metadata
Required format IDs
Available format IDs
Select rendering environment suitable for Digital Object
Rendering environment

Figure 3: Characterization Workflow

As a first step, the digital object's container format has to be determined and its internal structure or filesystem has to be accessed. In our case, all containers were CD-ROM images, but some were, unfortunately, not standard ISO 9660 images. Many CD-ROMs of the multimedia age were produced as *hybrid* CD-ROMs containing both, an ISO 9660 file system as well as one (or more) additional file systems to overcome the restrictions of ISO 9660 and to implement system-specific features.

A common example are HFS/HFS+ hybrid CD-ROMs produced to be used with Apple computer systems and IBM-PCs. While Windows or DOS only sees the ISO 9660 filesystem, MacOS was able to access the HFS filesystem contained on the same CD-ROM to provide a different view on the contents. Fig. 1 shows the content of a HFS-hybrid CD-ROM opened on an Apple Macintosh computer system while Fig. 2 shows the content of the same CD-ROM opened on a Windows computer system.

For characterizing the container's content, both "views" of an object need to be evaluated. Unfortunately, none of the tools currently included in FITS are able to characterize the container format correctly. Running FITS on a hybrid disk image file will return a `SINGLE_RESULT` classifying the file as `application/x-iso9660-image`. To correctly identify the HFS filesystem on these images, we had to implement an additional container classification tool, matching for `Apple_partition_map` and HFS/HFS+ Master Directory Block signature to detect HFS/HFS+ filesystems on the CD-ROM images.

Once the filesystems contained in the images are identified, characterization tools can be used to determine the file formats of all files contained in the object. To automate the process of finding a suitable environment, we aggregate file format information of all files found. The resulting required format IDs can then be matched against list of import or native file formats in the software environments that are available from the image archive. In this step, certain formats may have to be prioritized over others. For example, a multimedia CD-ROM contains many picture files that belong to a multimedia application. While theses files can be viewed by any software environment that can view e.g. JPEG files, the multimedia application itself may only run on a PPC Apple MacOS.

In the final step, a very similar matching process is used to find a hardware environment that fulfills the requirements of the software environment. The resulting set of digital object, software and hardware environment is then the complete rendering environment and can be used by the EMiL emulation framework to re-enact the object.

5. RESULTS

In a first round we classified 69 CD-ROMs which were carefully selected to reflect the the diversity of the memory institution's collections. The goal was to have a wide representation of objects for different time periods and different types of objects. For instance, objects with more or less complex applications, interactive multimedia content especially with audio and video, 3D rendering, etc., but also objects made for different operating systems (Mac/Windows hybrid CDs) were chosen. The CDs were published between 1991 and 2009. For most of these CD-ROMs we had a transcript of their original system requirements. These requirements are used as ground truth for evaluating our classification. This information was not available for a few items, which meant that these CD-ROMs had to be tested manually.

Based on the analysis of executables found on each CD-ROM, we could determine from 66 CDs at least one suitable rendering environment selecting an appropriate operating system. 35 CDs were classified with multiple environments. While 11 CDs were hybrids which can either be run under a Windows or a Mac OS environment, we found 24 CDs which provided binaries for different Windows environments (e.g. legacy support for Windows 3.11 or MS-DOS). Even though final characterization results could be stored as metadata along with each object both characterization tools and available environments may be improved over time. On-demand characterization would provide the best user-experience, since newly added or improved environments can be considered. On a 4 CPU machine using 8 parallel threads, characterization of a single CD took less the 30 seconds, in most cases even less then 10 seconds. Only one object took more than 60 seconds to process, which was in fact a 4 Gb DVD.

For some CD-ROMs however, our simple file-by-file classification approach failed, e.g. because no executables were present on the CD-ROM. For instance, we have found a "Chinese Language Course", which contained more than 6000 files, however, most were encoded as HTML (fmt/96), JPEGs (fmt/41,43) and WAV (fmt/143) format. Figure 4 shows the format distribution as histogram. With no browser or other executable on the medium, a characterization solely on executable formats cannot determine a suitable environment. In this case, the aggregate file format information indicates that any environment with a Web browser installed would be suitable.

In a similar case, only PPT files were present on the CD, which would require a software environment containing an Office installation and a CD which only contained images. In the specific case of CD-ROMs, presence of an *autorun.inf* indicates the requirement for a Windows environment. Such information can be used for a secondary classification step, by defining sets of file formats typically associated with specific software setups or domain specific applications. Even if no suitable environment is found, the analysis of file format distribution can give useful hints about (additionally) required setups for a specific collection.

6. CONCLUSION & OUTLOOK

We presented a first step towards an automated reading-room access workflow for a large digital media collection. Our goal was to support users when accessing a CD-ROM from a memory institution's catalog and ideally render it instantly in a suitable emulated environment. To achieve this we have implemented a characterization tool for digital media containers accompanied with a technical framework and workflows. The characterization workflow successfully determined at least one suitable environment for 66 out of 69 objects based only on executable file formats and time signatures. Hence, for a vast majority of objects, a very simple heuristic can be applied to automate access.

Our evaluation, however, also shows the limitations of such a simple approach. For instance, if no executables are present, no classification of a basic rendering environment is possible. For more sophisticated environments, e.g. an environment for typical office workflows or specific engineering tasks, a thorough description of the available software and its supported file formats is necessary. The characterization workflow is then easily able to find an environment that can

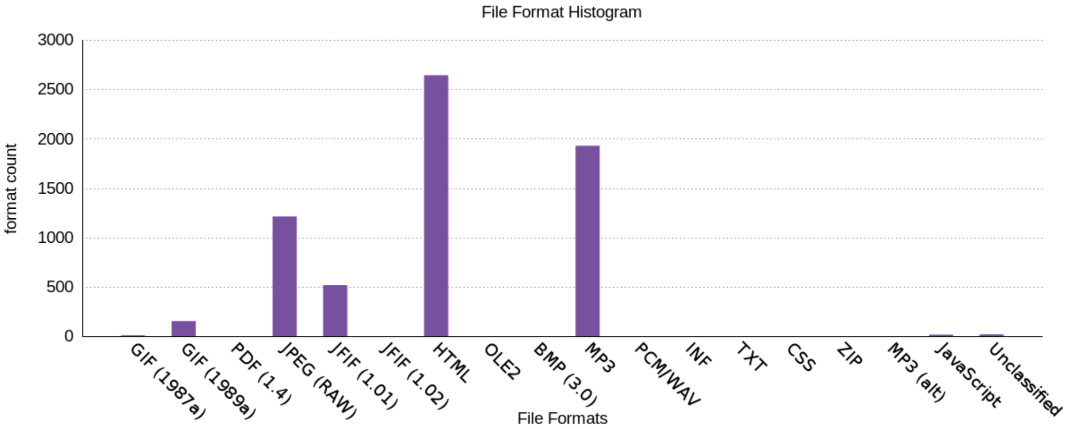


Figure 4: File format distribution of a CD-ROM containing an HTML-based language course.

render a specific file found on the CD-ROM.

For the aforementioned HTML-based language course, however, more considerations have to be taken into account. While almost any editor is able to open HTML files, the results may not be what the user expects. Similarly, all the WAV files found on the medium can be opened using a simple audio player. The object itself, however, has a characteristic ”Web medium” footprint and clearly should be opened in a Web browser that allows using the language course in the guided and interactive fashion that was intended, rather than opening every file individually. Therefore, a more complex characterization approach that recognizes a ”Web medium” footprint may be required.

A future option is to encode our technical metadata using the PREMIS data dictionary or similar established metadata standards to support interoperability and to simplify adaption or integration of emulation-based preservation strategies. In a similar way, technical interfaces to technical registries, such as PRONOM and TOTEM, enable rapid publication and sharing of new file format signatures, and especially, verified relations between file formats and necessary rendering software.

7. REFERENCES

[1] D. Anderson, J. Delve, and D. Pinchbeck. Document describing metadata for the specified range of digital objects. KEEP Public Deliverable D3.2a (online).

[2] T. Bähr, M. Lindlar, K. Rechert, and T. Liebetraut. Functional Access to Electronic Media Collections using Emulation-as-a-Service. In *Proceedings of the 11th International Conference on Digital Preservation (iPres14)*, page 332. State Library of Victoria, 2014.

[3] A. Brown. Pronom 4 information model. 2005.

[4] G. Brown. Developing virtual cd-rom collections: The voyager company publications. *International Journal of Digital Curation*, 7(2):3–22, 2012.

[5] 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. *IPRES 2014 proceedings*, 2014.

[6] A. Dappert, S. Peyrard, C. C. Chou, and J. Delve. Describing and preserving digital object environments. *New Review of Information Networking*, 18(2):106–173, 2013.

[7] A. Dappert, S. Peyrard, J. Delve, and C. C. Chou. Describing digital object environments in premis. In *9th International Conference on Preservation of Digital Objects (iPRES2012)*, pages 69–76. University of Toronto, 2012.

[8] J. Delve and D. Anderson. *The Trustworthy Online Technical Environment Metadata Database – TOTEM*. Number 4 in Kölner Beiträge zu einer geisteswissenschaftlichen Fachinformatik. Verlag Dr. Kovač, Hamburg, 2012.

[9] 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.

[10] T. Liebetraut, K. Rechert, I. Valizada, K. Meier, and D. von Suchodoloetz. Emulation-as-a-Service – The Past in the Cloud. In *7th IEEE International Conference on Cloud Computing (IEEE CLOUD)*, pages 906 – 913, 2014.

[11] D. Pinchbeck, D. Anderson, J. Delve, G. Alemu, A. Ciuffreda, and A. Lange. Emulation as a strategy for the preservation of games: the keep project. In *DiGRA 2009 – Breaking New Ground: Innovation in Games, Play, Practice and Theory*, 2009.

[12] PREMIS Editorial Committee. PREMIS data dictionary for preservation metadata, version 2.0. 2008.

[13] K. Rechert, D. von Suchodoloetz, T. Liebetraut, D. de Fries, and T. Steinke. Design and Development of an emulation-driven Access System for Reading Rooms. In *Archiving 2014*, pages 123–132. IS&T, 2014.

[14] J. van der Hoeven and D. von Suchodoletz. Emulation: From digital artefact to remotely rendered environments. *International Journal of Digital Curation*, 4(3), 2009.

[15] K. Woods and G. Brown. Assisted emulation for legacy executables. *International Journal of Digital Curation*, 5(1), 2010.

Human and Machine-Based File Format Endangerment Notification and Recommender Systems Development

Heather Ryan
University of Denver
Denver, Colorado, USA
heather.m.ryan@du.edu

Roman Graf
Austrian Institute of Technology
Vienna, Austria
roman.graf@ait.ac.at

Sergiu Gordea
Austrian Institute of Technology
Vienna, Austria
sergiu.gordea@ait.ac.at

ABSTRACT

Effectively preserving access to digital content over time is dependent on availability of an appropriate IT infrastructure including access to appropriate rendering software and its requisite operating systems and hardware. The complexity of this task increases over time and with the size and heterogeneity of digital collections. Automating notifications on file format endangerment and decision recommendations can greatly improve preservation planning processes. This paper presents work in progress that contributes to the design and testing of an automated file format endangerment notification and recommendation system. This system's design is based on concepts explored in previous research, but it presents the novel application of statistically generated similarity profiles and machine-generated recommendations based on human expert input.

General Terms

Frameworks for digital preservation; Preservation strategies and workflows

Keywords

File format endangerment, institutional risk profiles, recommender systems, notification systems

1. INTRODUCTION

Preserving access to content encoded in particular digital file formats requires the availability of the appropriate software and hardware infrastructure. Over time, it becomes incrementally difficult to maintain this particular infrastructure and to access the stored digital content (i.e. the hardware and/or software may reach their life end). For those managing large, heterogeneous digital collections, the challenge grows with the size and variety of content aggregated in their collections. This is particularly challenging for state and government archives, which are required to preserve all content produced by their supported government agencies, regardless of format. Web archives also pose unique challenges in preservation in terms of scale and complexity.

Knowing when certain file formats are becoming endangered, meaning in danger of becoming inaccessible using commodity hardware and software; and receiving recommendations for how to maintain access to the endangered format is an important component of a sound digital preservation workflow. Having these services augmented by expert opinion and semi-automated through appropriate software support can reduce the difficulty of this challenge.

Evidence collected through an interview-based study [1] and through personal conversations with individuals managing or working with digital collections in memory institutions suggests that there is a need for systematic file format endangerment measurement, notification, and recommendation. Some indicate that such efforts are not necessary [2][3]. Arguments against these efforts cite the inherent difficulty in quantifying many file format endangerment factors and general lack of trust in automated recommender systems as inhibitors to successfully measuring file format endangerment and providing alerts and recommendations for file format risks. Other underlying concerns around developing file format endangerment measures and tools lie in expert systems' apparent circumvention of individual expertise and lack of observable data to test these measures and systems.

To reconcile these concerns, we have extended the design of the File Format Metadata Aggregator (FFMA) [4]. It now includes: 1) expert informed, hybrid decision support tools, 2) a case-based recommender system that produces recommendations according to similarity metrics [5] and initial tests on a hybrid collaborative filtering system for building/identifying institutional profiles, and 3) a knowledge based system for computing the risk factors and levels [6] on test data collected for a previous file format endangerment study [7].

The present system requires additional evaluation and testing, both through testing the system components and algorithms, and through analysis of user needs and trust in automated systems. Our first goal is to collect data on which factors digital collection managers use to assess institutional preservation friendliness. Here preservation friendliness is related to a file format's various attributes that may contribute to or hinder preservability of digital content within an institutional context. Traditionally, it has been preservation friendly formats that are selected for inclusion in digital collections managed by memory institutions [8][9][10][11][12][13]. This data will be used as corpora for testing algorithms designed to calculate institutional risk profiles. Additionally, we aim to collect information on which file formats most commonly appear in study participant collections. We use the collected list of file formats, which are sufficiently documented in Linked Open Data (LOD) repositories such as DBPedia and Freebase, as a basis for further system testing.

Our second goal is to collect information on perceived trust and utility of an automated file format endangerment notification and recommender system. Issues of trust are common complicating factors in the design and implementation of recommender systems and it is important to address them early in system design. We use information collected from this portion of the study to inform the development of additional trust-building measures such as trustworthiness or transparency, and the ability to allow users to indicate or correct information [5].

This work in progress will lead to the novel approach to decision support for digital collection managers. While the system makes use of data-mining and statistical analysis of endangerment factors, it complements the machine learning aspects of the system with human expert input and recommendations.

This paper is structured as follows: Section 2 provides an overview of related work as well as existing work associated with this project, Section 3 explains the motivation behind each aspect of the study, Section 4 describes the study design and how it has to be applied to the design and testing of the file format endangerment notification and recommender system, and Section 4 concludes the paper and outlines planned future work.

2. RELATED WORK

This research builds on previous work on other similar efforts as well as our own related work. Similar initiatives PANIC [14], AONS II [15], SPOT [16], and the P2 Registry [17] incorporate file format identification and risk notification mechanisms.

A preliminary study has been conducted to assess file format endangerment factors [7] for measurability and fit for inclusion in a file format endangerment index. Once validated, the index will provide the framework for file format endangerment warnings. Algorithms and visualization components have been tested for risk profile definition [18] and format coherences [19].

This research improves on initial projects to extract file format data from various online resources [20] and to provide decision support using fuzzy logic [21]. Additionally, it pulls from earlier work on developing a File Format Migration Center that facilitates user-generated ratings and recommendations for file format conversion pathways [22]. The work in progress presented here builds on these previous efforts while developing novel technologies for data collection, data analysis, data visualization, alerts, and recommendation.

3. STUDY MOTIVATION

The current study is designed to contribute to the testing and further design of an automated file format endangerment notification system. Data collected is meant to be used as test corpora and to inform additional design decisions.

Initial tests of Naive Bayes analysis were performed using initial data collected for [7] which produced a successful proof of concept model for automated institutional risk profile generation. Sparse data and minimal ordinal values limited the degree to which this data set could be used for more robust testing.

Institutional risk profiles are created using human generated preference settings of institutionally-based file format evaluation factors. Recommendations for decision-making are made based on similarity calculations between the individual risk profile preferences. Similar institutional risk profiles will receive similar decision recommendations, based on expert input. It is necessary to collect more thorough input on file format evaluation factor preferences to accurately calculate the institutional risk profiles.

Previous tests of this and similar systems involved test file formats that were selected based on various criteria that may not be directly related to actual use-cases. Our intent with future system development is to test using file formats that are known to reside in digital collections currently managed in real institutional settings. To accomplish this, we are collecting information on the most commonly occurring file formats in collections managed by study participants.

Trust is a concern in the development of recommender systems, both trust in the other human contributors to the system and trust in the system's automated recommendations [7]. There are methods that can be used to ameliorate lack of trust, but presence of distrust must first be established before additional probes can be used to determine underlying reasons for extant distrust. Once reasons for user distrust are established, action can be taken to address the reasons within the design of the system.

4. STUDY DESIGN

The following study design reflects the needs outlined in the study motivation section. The study consists of an online survey administered using the Qualtrics online survey software.

4.1 Research Questions

This study is designed to answer the following research questions:

RQ1: Which factors do individuals working in libraries and archives consider to be most important when evaluating file formats for inclusion in an institution's digital collection(s)?

RQ2: Which factors do individuals consider to be causes of file format endangerment?

RQ3: To what degree do individuals working in libraries and archives believe that a file format endangerment notification and recommender system will improve their work and their preservation related decisions?

RQ4: To what degree do individuals working in libraries and archives trust the concept of an automated file format endangerment warning and recommender system?

4.2 Participants

Study participants are individuals working in libraries and archives who make decisions about digital file formats in collections they oversee. They are recruited using emails to listservs, direct email contact, and through word of mouth.

4.3 Survey Design

The survey includes four sections: Demographics, Utility and Trust, File Format Factor Rating, and Common File Formats. The study is comprised of six questions, where Question 4 contains 31 sub-questions, for a total of 36 items.

4.3.1 Section 1: Demographics

Q1. Institution type (e.g. Academic Library, City Archives, National Library, Medical Library, etc.)

4.3.2 Section 2: Utility and Trust

Q2. An application that is able to notify about endangered file formats and explain the nature of risks will improve my work and my preservations related decisions.

- Strongly agree
- Agree
- Neutral
- Disagree
- Strongly disagree

Q3. I trust a computer system which is able to indicate file formats that are in danger of not being supported by commodity hardware-software systems in the near future?(10-20 years).

- Strongly agree
- Agree
- Neutral

- Disagree
- Strongly disagree

Please explain your answer.

4.3.3 Section 3: File Format Factor Rating

Q4. **(**Please rate the following 31 factors based on how important they are to consider when selecting file formats that your particular institution is able to preserve access to in the near future (10-20 years):

- Not at all important
- Slightly important
- Moderately important
- Very important
- Extremely important

Factors:

1. **Availability Online** - the degree to which the format is available on the Web.
2. **Backward Compatibility** - whether or not newer versions of the rendering software can render files from older versions.
3. **Community Support** - the degree to which communities support the file format.
4. **Complexity** - relates to how much effort has to be put into rendering and understanding the contents of a particular file format.
5. **Compression** - whether or not, and the degree to which a file format supports compression,
6. **Cost** - The cost to maintain access to information encoded in a particular file format, e.g. to migrate files, to maintain the rendering software, or to run an emulation environment.
7. **Developer/Corporate Support** - whether or not the entity that created the original software that produces output in the file format continues to support it.
8. **Domain Specificity** - the degree to which the format is used only within specific domains.
9. **Ease of Identification** - the ease with which the file format can be identified.
10. **Ease of Validation** - the ease with which the file format can be validated, where validation is the process by which a file is checked for the degree to which it conforms to the format's specifications.
11. **Error-tolerance** - the degree to which this format is able to sustain bit corruption before it becomes unrenderable.
12. **Expertise Available** - the degree to which technological expertise is available to maintain the existence of software that can render files saved in this format.
13. **Forward Compatibility** - whether or not older versions of rendering software can render files from newer versions.
14. **Geographic Spread** - the way in which a file format is spread across the world; whether spread thinly across the globe or condensed heavily in a particular area.
15. **Institutional Policies** - the degree to which a file format is affected by institutional polices, such as whether or not an institutional policy states that content encoded in this format will be collected and preserved.
16. **Legal Restrictions** - the degree to which this file format is or can be restricted by legal strictures such as licensing, copy and intellectual property rights.
17. **Lifetime** - the length of time the file format has existed.
18. **Metadata Support** - whether or not the file format allows for the inclusion of metadata.
19. **Rendering Software Availabilty** - whether or not any type of software is available that can render the information stored in this file format.
20. **Rendering Software Functionality/Behavior Support** – the degree to which available rendering software supports various functionality and behavior encoded in a particular file format.
21. **Revision Rate** - the rate at which new versions of this file format's originating software are released.
22. **Specifications Available** - whether or not documentation is freely available that can be used to create or adapt software that can render information stored in this file format.
23. **Specification Quality** - (sub-factor of "Specifications Available") the understandability and usefulness of the format's available specifications in maintaining access to content encoded in that format.
24. **Standardization** - whether or not this file format is recognized as a standard for use and/or preservation by a reputable standards body.
25. **Storage Space** - the average amount storage space a file saved in this format requires when saved.
26. **Technical Dependencies** - the degree to which this file format depends on specific software (beyond typical rendering software), operating systems, and hardware in order for its contents to be successfully accessed or rendered.
27. **Technical Protection Mechanism** - whether or not this file format allows for or is encumbered by technical protection mechanisms such as Digital Restrictions Management (DRM).
28. **Third Party Support** - the degree to which parties beyond the original software producers support the file format.
29. **Ubiquity** - the degree to which use of this file format is widespread and in common use.
30. **Value** - the degree to which information encoded in this format is valued.
31. **Viruses** - the degree to which the format is susceptible to containing or being damaged by viruses.

The list of factors will be presented to participants in random order to enhance reliability of responses.

Q5. Which of the following factors [Backward Compatibility, Community Support, Complexity, Cost, Developer/Corporate Support, Expertise Available, Forward Compatibility, Legal Restrictions, Rendering Software Availability, Rendering Software Functionality/Behavior Support, Specifications Available, Specification Quality, Standardization, Technical Dependencies, Third Party Support, Ubiquity] do you believe is a/are direct cause(s) of file format endangerment (versus factors for evaluating whether or not a format is included in a preserved collection)?

4.3.4 Section 3: Common File Formats

Q6. Please list the most commonly appearing file formats in your institution's digital collection(s). For each file format listed:

Describe briefly their application(s) (e.g. historical photographs, institutional documents, medical records, GIS data, etc).

Explain briefly why the file format was selected for inclusion in your institution's collection(s). What advantages does it present over other, similar file formats.

5. CONCLUSION AND FUTURE WORK

The goals of this study are to inform the development of a semi-automated file format endangerment warning and recommendation system. The survey will provide insight into what participants think are the most important factors that individuals consider when evaluating file formats for inclusion in their collections. This data will serve as the test corpora for statistically determining institutional risk profiles, which will then be used to establish likeness between users. The study will also provide a use case based list of file formats that will provide a basis for realistic system experiments and tests. Lastly, the survey will help to establish the usefulness of an automated file format endangerment warning and recommender system, and to what degree people think they can trust and rely on an automated system.

Continuing research involves continued experiments and system tests, further examination of trust in automated recommender systems, and development of additional framework for system deployment and use.

6. REFERENCES

[1] Bowden, H. 2010. Assessing need for file for an automated file format obsolescence warning system for digital collections. In *iConference 2010* (Urbana-Champaign, IL, February 03 - 06, 2010).

[2] van der Knijff, J. 2013a, September 30. *Assessing file format risks: searching for Bigfoot?* Message posted to Open Planets Foundation blogs at http://www.openplanetsfoundation.org/blogs/2013-09-30-assessing-file-format-risks-searching-bigfoot

[3] van der Knijff, J. 2013b, October 8. *Measuring Bigfoot*. Message posted to Open Planets Foundation blogs http://www.openplanetsfoundation.org/blogs/2013-10-08-measuring-bigfoot

[4] Graf, R., and Gordea, S. 2012. Aggregating a knowledge base of file formats from linked open data. In *Proceedings of the 9th International Conference on Preservation of Digital Objects*, (Toronto, Canada, October 01- 05, 2012), 292–293.

[5] Ricci, F., Rokach, L., & Shapira, B. 2011. Recommender systems handbook. Springer, New York. DOI=10.1007/978-0-387-85820-3

[6] D. Heckerman. 1997. Bayesian networks for data mining. *Data Mining and Knowledge Discovery*, 1,1,79–119.

[7] Ryan, H. 2014. Occam's Razor and file format endangerment factors. In *Proceedings of the 11th International Conference on Digital Preservation,* (Melbourne, Australia, October 6-10, 2014).

[8] Arms, C.R., & Fleischhauer, C. 2005. Digital formats: Factors for sustainability, functionality, and quality. *Imaging Science & Technology Archiving 2005*, Washington, DC, (April 2005), 222-227

[9] Lawrence, G.W., Kehoe, W.R., Rieger, O.Y., Walters, W.H., & Kenney, A.R. 2000. *Risk management of digital information: A File format investigation*. Washington, DC: Council on Library and Information Resources.

[10] Huc, C., et al. 2004. *Criteria for evaluating data formats in terms of their suitability for ensuring information long term preservation*. Technical Report. Groupe Pérennisation des Informations Numériques.

[11]Cornwell Management Consultants. 2005. *Selection of preservation formats: trends and issues*. Technical Report. The National Archives, U.K.

[12] InterPARES. 2007. General study 11 final report: Selecting digital file formats for long-term preservation (Version 1.1). British Columbia, Canada: McLellan, E. P.

[13] Rog, J., & Wijk, C, van. 2008. *Evaluating file formats for long-term preservation*. Technical Report. Koninklijke Bibliotheek.

[14] Hunter, J. & Choudhury, S. 2006. PANIC: An integrated approach to the preservation of composite digital objects using Semantic Web services. *International Journal on Digital Libraries, 6*(2), 174-183.

[15] Pearson, D., & Webb, C. (2008). Defining file format obsolescence: A risky journey. *International Journal of Digital Curation, 3*(1), 89-106.

[16] S. Vermaaten, B. Lavoie, and P. Caplan. 2012. Identifying threats to successful digital preservation: the spot model rsik assessment. *D-Lib Magazine*, 18, 9/10 .

[17] Carr, L., Tarrant, D., Hitchcock, S. 2011. Where the semantic web and Web 2.0 meet format risk management: P2 Registry. *International Journal of Digital Curation, 6*,1, 165-182.

[18] Graf, R., Gordea, S., Ryan, H. 2015. A Bayesian classification system for facilitating an institutional risk profile definition. In *Proceedings of the 17th International Conference on Information Technology and Engineering (ICITE*) (Oslo, Norway, July 17-18, 2015).

[19] Graf, R., Gordea, S., Ryan, H. 2015. A tool for visualization and analysis of file format coherences. In *Proceedings of the 4th International Conference of Asian Special Libraries (ICoASL)*. (Seoul, Korea, April 22-24, 2015).

[20] 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, (Lisboa, Portugal, September 2-5, 2013)*.

[21] Graf, R., Gordea, S., Ryan, H. 2014. A model for format endangerment analysis using fuzzy logic. In *Proceedings of the 11th International Conference on Digital Preservation, (Melbourne, Australia, October 6-10, 2014)*.

[22] Bowden, H. 2009. File format migration center: Final project paper. http://longtermdata.com/docs/HBowden_ProjectPaper.pdf

Getting to the Bottom Line: 20 Digital Preservation Cost Questions

Matt Schultz
Grand Valley State University
1 Campus Drive
Allendale, MI 49401
1-616-331-5072
schultzm@gvsu.edu

Aaron Trehub
Auburn University Libraries
231 Mell Street
Auburn, Alabama 36849
1-334-750-1695
trehuaj@auburn.edu

Katherine Skinner
Educopia Institute
1230 Peachtree St, Suite 1900
Atlanta, GA 30309
1-404-783-2534
katherine@educopia.org

ABSTRACT

Getting to the Bottom Line: 20 Cost Questions for Digital Preservation is a cost-gathering resource created by the Outreach Committee of the MetaArchive Cooperative in Spring 2015. Launched during an Association of Southeastern Research Libraries (ASERL) webinar (https://vimeo.com/121926212) on March 11, 2015, this resource has been shared broadly with libraries, archives, and other institutions that have an interest in procuring digital preservation services. The easy-to-use resource is designed to equip institutions with questions that they can use to identify the full range of costs that might be associated with any particular digital preservation service--proprietary, community-sourced, or otherwise. For a variety of reasons, services of all types do not always make their costs as transparent as institutions might prefer. Using the *Getting to the Bottom Line* question-set will help ensure that institutions do not leave any stones unturned when evaluating their options and that they gather the information that they need to make informed choices that lead to sustainable solutions. Institutions are encouraged to make free use of the questions, adapt them as needed, and provide feedback on their usefulness. Going forward, the resource will serve as a foundation for building additional and more sophisticated cost transparency resources targeted toward the digital preservation community.

General Terms

Institutional opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows; Training and education.

Keywords

Economics. Digital Preservation Costs.

1. INTRODUCTION

The question of cost—and the need for good cost models—has received extensive attention in the literature on digital preservation, including the 1996 Garrett-Waters report on “Preserving Digital Information”, the 2010 final report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (“BRTF-SDPA”), and the published volume *Aligning National Approaches to Digital Preservation* (Educopia, 2012). There has been no shortage of cost models for digital preservation, including (in roughly chronological order) products of the LIFE 1, 2, and 3 projects (UK: 2006-2012), Charles Beagrie’s Keeping Research Data Safe (KRDS) projects (UK: 2008-2011), the Cost Model for Digital Preservation (CMDP) project (Denmark: 2009-2012), the APARSEN/APA project (EU: 2010-present), the California Digital Library-Total Cost of Preservation (CDL-TCP) project (USA: 2012-present), and the 4C project (EU: 2013-2015), not to mention the work of David Rosenthal (LOCKSS), Serge Goldstein and Mark Ratliff (the Pay-Once-Store-Forever formula), Adrian Brown (“estimated value of digital assets”), and others.

This work has proven influential at-and-beyond the field level, and some of it has produced tools that people are using today—for example, the 4C Project’s Web-based Curation Costs Exchange (CCEx) calculator. There is, however, a lack of comprehensive, comparable, and reliable cost and pricing information on the digital preservation solutions—commercial, quasi-commercial, and community-based—that have emerged in the past decade and that are vying for the attention (and the money) of decision-makers at cultural memory institutions, media companies, and government departments. These decision-makers are interested in more than digital preservation theory. They want to know what a given solution is going to cost, and how that cost compares with the costs, both up-front and “hidden”, of other solutions in the marketplace.

Getting to the Bottom Line: 20 Cost Questions for Digital Preservation builds upon an “Action Session” on this topic at the Aligning National Approaches to Digital Preservation II (ANADP II) conference in Barcelona, Spain in November 2013 and, more recently, a webinar given in March 2015 as part of an ongoing series of webinars on issues in digital preservation sponsored and hosted by the Association of Southeastern Research Libraries (ASERL). As we will discuss, this list of twenty cost questions for digital preservation solutions compiled by the MetaArchive Cooperative Outreach Committee in late 2014-early 2015 provides a much-needed basis for practical cross-comparison between digital preservation solutions. Herein, we begin by providing a brief overview of transparency problems in digital preservation. We then describe the list of questions, as well as the initial results of applying them against extant solutions and the implications thereof.

2. The Problem of Cost Transparency

Digital preservation and digital archiving services and solutions are becoming widely available. Services and solutions traverse a spectrum from the strictly commercial (e.g. Amazon Glacier, Preservica, Google Cloud Storage Near Line, etc.) to the community-based (e.g., LOCKSS, MetaArchive Cooperative, Alabama Digital Preservation Network, etc.). There are also services that fall somewhere between the two ends of this spectrum, representing a mixture of community-developed technologies that are then hosted and offered from within commercial environments such as Amazon or Rackspace (e.g. DuraCloud, ArchivesDirect, etc.).

Each service/solution offers its own unique technical approach to preserving digital information and does so via very different business models. Institutions have, for the first time, a range of maturing options that can help them to address the challenge of preserving their unique digital assets. However, understanding the significant differences between these services is challenging, even for seasoned professionals (see e.g., the POWRR Project). Establishing a clear understanding of what features each service offers, how those services might fit together to inform a preservation workflow, and what costs will be associated with each service *and* with bridging services together is daunting at best, and nearly impossible to project at worst.

Understanding at a distance the range of costs that might be associated with any given service/solution is critical in the early stages of evaluating options. For a variety of reasons, this information is not always easy to obtain. Service providers may not have a fixed pricing schedule and instead prefer to negotiate pricing individually with customers. Still others withhold pricing information when it does not place them at a competitive advantage for a particular service offering (e.g., storage or subscription/licensing/membership fees). And finally, many services withhold pricing information in order to ensure that prospective customers will speak directly to them about their service offerings rather than relying on a cost sheet alone. These are all very common and familiar barriers to cost transparency, and they are encountered more generally in the library services marketplace.. A culture of silence has permeated many service offerings—database subscriptions and scholarly journals pricing to note a couple of prominent examples, where confidentiality clauses have helped to occlude differences in costs assessed against various institutions for access to the same content (see e.g. Bergstrom, McAfee, and Courant's work in this area).

As academic research libraries and data centers, public libraries, historical societies, museums, and other scholarly/cultural institutes seek digital preservation services/solutions, they must consciously demand cost information, and they must use that information to critically evaluate both the services available and the costs of those services. The *20 Cost Questions* are intended to empower institutions; helping them to gain the information they need regarding a range of digital preservation costs. Our hope is that their use within the community will help the entire field to avoid the longstanding transparency problems we have faced in other areas of service procurement.

Transparency in costs for digital preservation--ensuring that institutions can make sustainable choices and avoid hidden costs that might undermine their preservation missions--is vitally important as this field continues to mature and more and more services become available. Admittedly, cost transparency is often viewed as a risk factor for service/solution providers. However, when the full relationship between costs and service offerings is made more widely available it provides each service/solution provider with important information about how their offerings compare to those of others, and data that can be used for improvements and standardization of both services and business models. Furthermore, transparency around costs does not have to equate to a race to the bottom when it comes to pricing, rather it is an opportunity for a service/solution provider to clearly argue for the excellence and return on investment of their unique approach to solving the challenges of digital preservation.

3. Tackling the Transparency Problem

Getting to the Bottom Line: 20 Cost Questions for Digital Preservation includes the diverse perspectives of academic and public library representatives who met over the course of several months to actively discuss the barriers to transparent cost gathering that they have experienced. The libraries involved in its development included Auburn University, Greene County Public Library, Indiana State University, Purdue University, and the University of Tennessee, Knoxville. Each of these libraries has been an early adopter of digital preservation services and solutions. In some cases, these libraries have chosen to experiment with and/or use multiple service offerings for the sake of comparison and benchmarking. All have ample anecdotal and evidential information from several years worth of their own efforts to advance their digital preservation agendas. In addition to identifying common barriers, they also clearly delineated the full range of digital preservation activities to which services/solutions tend to assign fees. As these institutions worked together to craft questions that other institutions could use to navigate the waters of cost transparency, they made intentional efforts to incorporate the concerns of smaller, under-resourced institutions. Through structured dialogues and interviews with smaller institutions, concerns around sustainability, requirements for local expertise, and availability of support services, among other concerns, were emphasized and given proper recognition via the questions.

Below are some examples of the 20 questions we encourage organizations to present to prospective solution providers.

1. What are the solution provider's licensing, subscription or membership fees?

   -Have these fees increased or decreased over the past three years, and why?

   -How often is the fee structure reviewed? And how are fees set?

   -How are customers/subscribers/members consulted during any such reviews?

2. Is there a minimum licensing/subscription/membership term?

3. On average, how long does it take to begin using the solution once a contract or service license agreement (SLA) has been signed?

   . -What steps are involved?

4. In terms of sustainability, does the solution provider have a strategic plan, succession plan, or disaster recovery plan?

   . -If so, how up-to-date are such plans?

   -Has the solution provider engaged in any audits or risk assessments?

   -Are any of the plans or audit/assessment results publicly available?

The full set of cost questions is available here: http://www.metaarchive.org/cost-questions.

4. MOVING FORWARD

The *Getting to the Bottom Line* question set was published in early 2015 and has already gained interest and currency. The MetaArchive and its extended community of like-missioned institutions look forward to gathering further feedback on the questions and taking this timely work on cost transparency to the next stage: namely, a Web-based matrix for collecting and comparing costs for various digital preservation solutions, using an agreed-upon set of cost elements derived from the question set and community feedback. We propose that the matrix be hosted and maintained at a community-driven and oriented organization, of which there are several respected candidates. For example, the Open Preservation Foundation (http://openpreservation.org/), the Digital Preservation Coalition (DPC), or the Community Owned Digital Preservation Tool Registry (COPTR: http://coptr.digipres.org/Main_Page) to name just a few. A stable organizational host can help to ensure that the resource is actively used and maintained by both the digital preservation community as well as solution/service providers from whom cost information would need to be solicited.

5. ACKNOWLEDGMENTS

Our thanks to the MetaArchive Outreach Committee for their efforts to improve transparency in the digital preservation environment.

6. REFERENCES

[1] 4C project (EU: 2013-2015). http://4cproject.net

[2] APARSEN/APA project (EU: 2010-present). <http://www.alliancepermanentaccess.org/index.php/knowledge-base/digital-preservation-business-models/costbenefit-data-collection-and-modelling>

[3] Beagrie, Neil, Brian Lavoie, and Matthew Woollard (2010), *Keeping Research Data Safe 2* <http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf>

[4] Bergstrom, Ted; McAfee, R. Preston; and Courant, Paul. Big Deal Contract Project. <http://www.econ.ucsb.edu/~tedb/Journals/BundleContracts.html>

[5] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. February 2010. <http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf>

[6] Brown, Adrian. "Cost Modeling: The TNA Experience." The National Archives (UK). PowerPoint slides presented at the DCC/DPC joint Workshop on Cost Models, held July 26, 2005. <http://www.dpconline.org/docs/events/050726brown.pdf>

[7] California Digital Library-Total Cost of Preservation (CDL-TCP) project (USA: 2012-present). <https://wiki.ucop.edu/display/Curation/Cost+Modeling>

[8] Cost Model for Digital Preservation (CMDP) project (Denmark: 2009-2012). <http://www.costmodelfordigitalpreservation.dk>

[9] Goldstein, Serge J., and Mark Ratliff (2010), *DataSpace: A Funding and Operational Model for Long-Term Preservation and Sharing of Research Data*, <http://arks.princeton.edu/ark:/88435/dsp01w6634361k>

[10] LIFE 1, 2, and 3 projects (UK: 2006-2012). <http://www.life.ac.uk/>

[11] Rosenthal, David (2011), "Modeling the economics of long-term storage," *dshr's blog*, September 27, 2011 <http://blog.dshr.org/2011/09/modeling-economics-of-long-termstorage.html>

[12] Wheatley,Paul, P. Bright, and Rory McLeod (2008), *LIFE Generic Preservation Model v.1.4* <http://discovery.ucl.ac.uk/14128/1/14128.xls>

Best until … A National Infrastructure for Digital Preservation in the Netherlands

Barbara Sierman
KB, National Library of the Netherlands
PO Box 90407
2509 LK The Hague
+31 70 314 01 09
Barbara.Sierman@KB.nl

Marcel Ras
NCDD, National Coalition for Digital Preservation
PO Box 90407
2509 LK The Hague
+31 6 147 77 671
Marcel.Ras@ncdd.nl

ABSTRACT

This paper describes the developments in the Netherlands to establish a national Network for Digital Heritage. This network is based on three pillars: to make the digital heritage visible, usable and sustainably preserved. Three working programmes will have 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.

General Terms

Infrastructure opportunities and challenges

Keywords

Digital Preservation, NCDD, national infrastructure

1. INTRODUCTION

From the very beginning, “collaboration in digital preservation” was a phrase used by many professionals in the field, “as no one institution can do digital preservation on its own”. Voluntary collaboration between countries or partners in the same domain (libraries, data centers) found a more firm implementation in organizations like nestor (2003) [1], DPC (2002) [2], NDSA (2010)[3] and NCDD (2007)[1], often serving as a platform for training, knowledge exchange and the study of specific preservation-related issues.

With the growing amount of digital material and organizations involved in preserving it, to foster more collaboration between these organizations by establishing an infrastructure on a national level presents itself as a logical next step. The level of maturity in digital preservation, where now some basic principles are established, will also contribute to this development.

Some examples were already presented. At last year’s iPRES the national digital repository of Ireland was discussed, that will host digital collections of a variety of Irish institutes [4]. Darryl Mead from the National Library of Scotland described the effort to create a national preservation infrastructure in Scotland [5]. And in Finland the national library, archives and museums already share an infrastructure. [6] 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 realisation of further cost reductions by improving efficiency of the workflows necessary to undertake digital curation” [7].

The Netherlands is no exception in this pattern and is now taking steps to create a national infrastructure for digital preservation.

1.1 The Dutch digital landscape

The term “national digital heritage” does not only cover the collections of the so-called cultural organisations, like archives libraries and museums. It also covers the scientific heritage, as collected by universities, research organisations and data centres. Therefore, the national digital heritage of the Netherlands is not collected in one place but is preserved by a set of national organisations together. Records of the public broadcasting, for example, are preserved by the Institute of Sound and Vision. Research data in the humanities and social sciences are preserved by Data Archiving and Networked Services (DANS) while research data from the technical universities are collected and preserved by 3TU, a collaboration of 3 technical universities. The National Library of the Netherlands is mandated by law to focus on the preservation of “publications” (without an explicit legal deposit law prescribing what should be part of the collection). The National Archive is responsible for governmental archives.

The organisations acted independently in the past but changing requirements in the digital age, both from a point of view of the users of digital material as well as from an efficiency perspective,

[1] 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) which guarantees long-term access to digital information in The Netherlands.

will require some stronger collaboration and a clear description of roles and responsibilities.

According to figures from the Enumerate data platform for the Netherlands [8], the national collection consists of 44 million museum objects, 845 kilometre of archives, 9 million publications and 1.5 million hours in audio-visual collections.

1.2 Access to the Dutch digital heritage

More and more collections belonging to archives, libraries, media, museums, and knowledge institutes are being digitised and made available online. Institutions are developing functional and technological facilities for making these collections available for digital access and use, simultaneously making the maintenance of these collections cost-effective and sustainable. These are often comprehensive programmes unique to the logic, solutions, and dilemmas that are common in that particular domain.

These are exciting times for archives, libraries, and museums. They are realising that, in the information society, their collections are goldmines. At the same time, the digital environment has made it impossible for them to continue overseeing the entire process of acquiring and managing their collections, and then making them available. For every work process, institutions are often using technology that is developed and managed by someone else.

Institutions that are charged with managing heritage collections and making them accessible are finding themselves in the position of having to redefine their roles. The questions they might ask themselves in this endeavour include:

- How do we reach new user groups? How do we engage them, and what services do we offer them?
- How can we carry out our mission while complying with copyright laws?
- What competencies do we need to be successful in a digital context?
- What are the costs and benefits of making collections available to the public?
- What facilities will we manage ourselves, what services will we purchase, and where will we link to other infrastructures?

The main challenge is to make the national digital heritage accessible for a wide range of users, anytime, anyplace. Developing sector-wide infrastructures and increasing their interconnection will help organisations to do so. By coordinating their IT strategies, parties can achieve benefits of scale and reuse existing building blocks. Making smart connections between collections will enable users to view, experience, and re-use each object in a much richer context. When it comes to digitisation, the major challenges facing the heritage sector relate to scaling up their facilities to be more effective and efficient and linking the collections together to facilitate use.

Being able to meet wide-ranging and constantly changing user demands will depend on having customised and flexible digital facilities. However, upscaling and standardisation are needed to lower costs, improve compatibility, and increase sustainability. This is why any facilities developed must be as reusable as possible.

2. NETWORK DIGITAL HERITAGE (NDE)

Initiated by the Ministry of Education, Culture and Science, the Network Digital Heritage (NDE) was set up in 2014. The participants in this network are national organizations with large digital collections and a mandate to preserve them, like the National Library (Koninklijke Bibliotheek or KB), the Institute of Sound and Vision (BenG), the Cultural Heritage Agency, the Royal Netherlands Academy of Arts and Sciences (KNAW), the National Archives (NA), the Cultural Heritage Agency of the Netherlands (RCE) together with other partners like, for example, the knowledge centre (DEN) and the National Coalition for Digital Preservation (NCDD). This Network Digital Heritage is a partnership that focuses on developing a system of national facilities and services for improving the visibility, usability, and sustainability of digital heritage.

The Network presented a National Strategy for digital Heritage in 2015 [9]. This strategy offers a perspective on developing a national, cross-sector infrastructure of digital heritage facilities. It contains objectives, starting points, and specific work programmes for a joint approach. The national strategy is the result of a one-year process. During this year dozens of professionals from the various sectors have contributed by engaging in working groups, attending meetings, and reviewing texts, including a public consultation.

A general principle is that no new facilities will be developed or new tools will be created, but that (in principal) existing facilities will be used or, if necessary adapted for better and broader use.

Another principle is that these efforts are focused on the user of this national heritage, now and in the future.

Implementing this strategy will require efforts at various levels. Individual institutions will develop an information policy and link their collections, knowledge, and facilities to a larger network. Assigning an active role to five sectorial organisations, so called “hubs”, will reinforce cooperation within sectors. The “hub organisations” have a track record in their domain and long-term sustainability. The “hub organisations” are: KB, BenG, NA, KNAW and RCE. They work within the network as a cross-domain partnership, open for other organisations to join. Commercial parties are explicitly not excluded from this network. Cooperation with industry organisations, user groups, governments, and international networks will be promoted.

The shared strategy must result in more facilities being connected, standardised, and jointly developed and managed in the coming years. This will require more cooperation and knowledge sharing between the various heritage sectors, governments, producers, knowledge institutions, intermediaries, and users. They are working on shared principles, standards, and new methods of knowledge sharing. Agreements and choices sometimes involve a degree of obligation to benefit interoperability or efficiency. This will ensure the development of an infrastructure that is helpful and stimulating for individuals, as well as for large and small institutions, businesses, and governments.

Cooperation will be based on existing sectorial facilities, responsibilities, and funding flows. Working from that foundation, the parties will seek out opportunities for linking and upscaling facilities, as well as for eliminating obstacles. A better understanding of user wishes, the need for a more efficient use of public funds, and the potential of the partnership will reinforce the parties’ readiness to change the existing situation.

Starting from existing facilities and services that have been established in recent years, the Network Digital Heritage defined three work programmes to put the shared strategy into practice. This should help to move from the current decentralized approach in which cultural heritage institutes organised preservation by themselves, towards a more shared approach. Clearly not a centralised approach as the Dutch government will not set up facilities on a must use basis. But helping and stimulating cultural heritage organisations to make use from (existing) facilities on the basis of sharing.

3. THREE WORK PROGRAMMES

These Work Programmes are initiated to realize the goals set in 2015-2016. Their goals are summarized in the slogan “Zichtbaar, Bruikbaar, Houdbaar”, translated as Making digital heritage visible (*Zichtbaar*), Making digital heritage usable (*Bruikbaar*) and Sustainable preservation of digital heritage (*Houdbaar*):

1. Work programme 1 (Visible): Making digital heritage visible (*Zichtbaar*). This should increase the visibility of collections, explore user demand, and promote the use and re-use of digital collections.
2. Work programme 2 (Usable): Making digital heritage usable (*Bruikbaar*): This should improve the possibilities for using collections by making them jointly accessible online, connecting and enriching data using lists of terms and thematic management, and developing targeted services.
3. Work programme 3 (Sustainable): Sustainable preservation of digital heritage (*Houdbaar*). For a preservationist this is the interesting part, although highly connected with the other working groups. The aim is to work on the cross-sector sharing, utilisation, and scaling up of facilities for sustainable preservation and access, while devoting attention to cost management and the division of duties. More details of this Work Programme 3 will be described in paragraph 5.

4. THE NATIONAL COALITION FOR DIGITAL PRESERVATION

The activities in this third work programme are based within the framework of the National Coalition for Digital Preservation, a partnership between the Dutch National Library, the Dutch National Archives, the Dutch Institute for Sound and Vision, Data Archive and Networked Services and several cultural heritage organisations. It is a member organisation, funded by the participating organisations above mentioned with additional funding from the Ministry of Education, Culture and Science.

The NCDD was established in 2008, as a national coalition designed to promote the preservation and the usability of digital materials comprising the cultural and scientific heritage of the Netherlands. NCDD is 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.

A national survey on the state of affairs in digital preservation carried out in 2009 [10] gave a better understanding of the then-present status of digital preservation in the Netherlands. According to the outcomes of this NCDD survey, problems could be best addressed by developing a distributed national network for managing digital resources in the public sector. This infrastructure was understood to include not just storage facilities, but also a whole range of less tangible matters: a clear definition of roles and responsibilities, selection criteria, quality criteria, shared services, knowledge and expertise. The network should be based on collaboration between stakeholders, because the resources required by long-term digital preservation exceed the means of most individual institutions.

Following on the national survey, the NCDD in 2010 formulated a strategic agenda. 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 2009 survey. 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 to 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.

In 2013 NCDD made it part of its strategy to work on this collaborative model that should result in a distributed national infrastructure [11]. The first results are becoming available now (spring 2015). A roadmap for certification of Dutch digital repositories has been shaped, workflows for ingest of various types of born-digital materials are described and a scenario for a distributed infrastructure for permanent access has been laid out. This national distributed infrastructure will be based on a reference model developed by the NCDD in which all elements as services are laid out. Services can be everything from storage to preservation watch. The basic starting point is that infrastructures are in place, services are developed and facilities are already shared. But these facilities need to be scaled up, standardised and offered to more and different organizations, sometimes in different domains. These “service seekers” should be enabled to find the best services for their needs and have the professional skills to make the right judgements.

The next steps will be worked out in Work Programme 3 of the NDE, where the current situation will be turned into a networked future.

The efforts of the partners in the NCDD have led to an important result, namely that the preservation issues are addressed on a governmental level, and will be addressed in the goals set in the National Strategy. NCDD is commissioned to lead Work Programme 3 on sustainable preservation in 2015 and 2016.

5. SUSTAINABLE DIGITAL HERITAGE

The objective of Work Programme 3 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 increased effectiveness, greater efficiency and cost reductions. As already described, some of the activities in this work programme have been started and scheduled within the NCDD.

The work towards this goal is being done along three lines:

1. Better utilisation and upscaling of facilities
2. Cost management
3. Roles and responsibilities in digital collection development.

5.1 Better utilisation and upscaling of facilities.

Two examples will be explained here in more detail: storage facilities and persistent identifiers. Apart from these, some other (smaller) projects will be started with regards to, for example, participating in investigations towards a software repository for tailor made software used in research or art projects, national collaboration in file format research and preservation watch.

Storage facility for permanent access.

This facility will be especially focused on small organisations, which currently have no or hardly any professional facility for permanent storage of digital material. Research showed that this is the case in several specialized cultural and research domains like digital photography, digital art, humanities and architecture. During the 2 years of the program, an inventory of existing storage facilities in cultural heritage organisations in the Netherlands will be created. Based on this list, a small set of representative organisations in the above mentioned domains will be connected to these existing facilities. Apart from making use of facilities of their colleagues in the cultural heritage sector, there will also be the opportunity to make use of commercial partners. A programme for training staff will increase their knowledge of digital preservation. This will make staff more professional and enable them to either manage their digital collections or to outsource this task. Suggested models of service level agreements for different levels of preservation will be designed in collaboration, so that staff will well understand the terms and consequences. This project has a strong connection with the projects in Roles and Responsibilities and Cost Management.

Persistent identifiers

Facilities for assigning persistent identifiers to digital objects need to be implemented and this is highly related to Work Programmes 1(Visible) and 2 (Usable). In close collaboration with existing organisations distributing persistent identifiers (like DataCite Netherlands, Institute of Sound and Vision, DANS and 3TU), existing facilities and their use will be inventoried and a nationwide model will be designed. The goal is to make these facilities affordable and within reach of small and medium cultural heritage organisations as a web service offered by professional providers. A sound overview of the related costs for this facility is needed to estimate operational budgets. There is no intention to restrict the use to a limited set of identifiers.

5.2 Cost management

With regards to costs of preservation, valuable work was done in the Collaboration to Clarify the Costs of Curation (4C) project.[2] The tool they developed, the Cost Exchange Tool (CCeX) will be used to collect cost figures from the main part of organisations in the Netherlands with a preservation mandate, especially the above mentioned "hubs". This will require a different way of reporting from the financial administrations. Within two years it is planned that the main players in the Netherlands will have entered their key figures in CCeX, based on which a benchmark is planned. Training and communications will help to prepare organisations and share experiences.

5.3 Roles and Responsibilities

The activities under this theme aim to achieve a better-integrated way of and collaboration around selection, maintenance and providing access to collections.

Current collection policies are often a continuation of existing collecting policies, established in a physical world. Building digital collections requires evaluating these traditional collection policies, as digital objects poses the boundaries of what, for example, a "publication" is. There is a serious risk that certain digital objects belonging to the Dutch national heritage are not collected at all and that other digital objects are collected by more than one organisation. Getting an overview of "who is collecting what" will lower the risks of gaps or duplicated preservation activities. One of the projects will be related to web archiving, which is done by various organisations in the Netherlands, currently without an overview of the results.

Also the interconnectedness of digital objects will require more streamlining of preservation policies between organisations. A few pilot projects will be undertaken to support a new way of thinking, like archiving Interactive Media Assets and preservation of "enhanced publications" or "digital objects in digital context".

Setting up a Dutch national infrastructure in which facilities will be shared and offered by various organisations, will require a certain level of openness and trust. It will become more important to be open about the various preservation approaches and tune in with other organisations. Sound and published preservation policies will contribute to this openness. Training in preservation planning and watch will be developed to support organisations in developing their own preservation policies. This work will be based on the results from the European project SCAPE, where a Catalogue of Preservation Policy Elements [12] was created, as well as an overview of existing Published Preservation Policies [13].

To establish trust in digital preservation, a set of certifications are available, combined in the European Framework [14], consisting of the basic level Data Seal of Approval, the self-assessment level of the German standard DIN 31644 and the highest level of ISO 16363, the TDR. In the Netherlands a new tool, the Scoremodel, [15] was developed by DEN especially for small cultural heritage organisations, which is a starting point.

Some organisations in the Netherlands already acquired a DSA certificate, like DANS, 3TU. But many of the larger organisations have not. A roadmap has been developed with the aim to get the larger organisations like the National Archive, the National Library and the Institute of Sound and Vision DSA certified before 2018, with other repositories soon to follow.

6. EXPECTED RESULTS AND BENEFITS

The presentation of a national strategy and the establishment of 3 Work Programmes 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. Exemplary was an expert meeting organised by the NCDD in February 2015 to discuss this proposed infrastructure. On this occasion over eighty Dutch preservationists (and some Belgian colleagues) came together and discussed the national plans, sharing approaches, plans and doubts. The level of knowledge exchange and the willingness to collaborate were promising and proofs that we have made important steps forward. It is a fair promise for the next steps to be taken.

7. REFERENCES

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

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

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

[2] http://4cproject.eu/

[4] 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

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

[6] 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

[7] Investing in Curation. A shared Path to Sustainability http://www.4cproject.eu/d5-1-draft-roadmap

[8] http://enumeratedataplatform.digibis.com/datasets

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

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

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

[12] http://wiki.opf-labs.org/display/SP/Catalogue+of+Preservation+Policy+Elements

[13] http://wiki.opf-labs.org/display/SP/Published+Preservation+Policies

[14] European Framework, see http://www.trusteddigitalrepository.eu/Welcome.html

[15] DEN Scoremodel Digitale Duurzaamheid http://www.den.nl/standaard/383/

Techniques for Preserving Scientific Software Executions: Preserve the Mess or Encourage Cleanliness?

Douglas Thain, Peter Ivie, and Haiyan Meng
Department of Computer Science and Engineering
University of Notre Dame
{dthain|pivie|hmeng}@nd.edu

ABSTRACT

An increasing amount of scientific work is performed *in silico*, such that the entire process of investigation, from experiment to publication, is performed by computer. Unfortunately, this has made the problem of scientific reproducibility even harder, due to the complexity and imprecision of specifying and recreating the computing environments needed to run a given piece of software. Here, we consider from a high level what techniques and technologies must be put in place to allow for the accurate preservation of the execution of software. We assume that there exists a suitable digital archive for storing digital objects; what is missing are frameworks for precisely specifying, assembling, and executing software with all of its dependencies. We discuss the fundamental problems of managing implicit dependencies and outline two broad approaches: preserving the mess, and encouraging cleanliness. We introduce three prototype tools for preserving software executions: Parrot, Umbrella, and Prune.

General Terms

Frameworks for digital preservation

Keywords

software preservation, dependency management

1. INTRODUCTION

While it has long been common for scientific publications to be prepared via computer, today much scientific work is now done completely from beginning to end in a computer. An elaborate model system may be run in simulation, generating raw data which is then processed by complex analysis software, which produces outputs that are displayed by visualization software, which can then be included in a final publication for dissemination and peer review.

Much early work in digital preservation from the library community focused on preserving the final artifact of that chain of effort: the publication. This includes accounting for physical media decay and obsolescence in addition to ensuring the availability of software for interpreting the data so it can be displayed to a user[24]. However, scientific productivity and integrity depends significantly upon our ability to preserve, share, and use the earlier steps in that chain, including both the software and the data. A peer-reviewer might wish to delve into the data associated with a paper, beyond the summary graph presented by the author. A collaborator might wish to pick up the current experimental software stack and adjust some parameters in order to obtain a new result. A competitor might wish to evaluate a completely new technique and compare it with a published technique in order to ensure that the previous technique has been validly recreated.

Unfortunately, the current state of the art is not encouraging. For example, in the biotech industry, Amgen attempted to reproduce 53 "landmark" articles in cancer research. They only succeeded with 10% of them [3]. In pharmaceuticals, Bayer was only able to reproduce about 21% of published results in 67 different projects [22]. Other efforts [25] have pointed out that there is a clear gap between preservation policies and practices. One can easily see why: a published computational result may briefly state that it ran with a certain version of software on a certain operating system, but may fail to state critical configuration values, dependent software, or even the precise inputs to the program. It is a common tale, even in the field of computer science, that an experiment was not published with enough details to accurately verify the results.

In this paper, we consider from a high level what techniques and technologies must be put in place to allow for the accurate preservation of the execution of software. We assume that there exists a suitable digital archive which can preserve digital objects for the long term, as are now commonly in place at university libraries, academic publishers, and so forth. The challenge lies in precisely identifying what must be preserved, naming each object appropriately, and providing a means for the consuming user to reassemble and verify the result.

The fundamental challenge throughout is the matter of **implicit dependencies.** In our current systems, it is all too easy for the user of a computer to consume some resource (a file, a program, a web site) without explicit knowledge that they are doing so. This leads us to two broad approaches to software preservation: **Preserving the mess** involves allowing the user to keep working in the current way while supplementary tools identify dependencies automatically. **Encouraging cleanliness** requires the user to state more clearly in advance what they are attempting to do. As we will show below, preserving the mess is easy but results in preserved objects that are of little use beyond identical verification, while encouraging cleanliness is harder but encourages extension and comparison.

Along the way, we give an overview of three pieces of software that demonstrate some of these approaches to software preservation. **Parrot** [16] enables the end user to preserve a mess by automatically capturing the file and network dependencies that form the environment of an application. **Umbrella** [17] encourages cleanliness by providing a precise way to specify and instantiate a software execution environment. **Prune** goes further by tracking and recording a software execution in the form of of individual operations that build upon each other's outputs. Each of these prototypes has been developed in the context of an NSF-supported project, called Data and Software Preservation for Open Science (DASPOS), which is examining the needs of preservation for the high energy physics community.[1]

2. SIMPLIFIED EXAMPLE

We consider the following simplified example in order to define some terms and highlight preservation challenges that we have encountered in working with a variety of applications. Suppose that a user has a laptop running GreenSock Linux 8.3 and wishes to run an open source simulation program `mysim` 3.2 on a custom input file `data` to produce a single output file `result` by typing the following command into the terminal:

```
$ mysim -in data -out result
```

The user's objective is now to preserve not just the software itself, but that specific *execution* of the software, so that others can verify a result and also extend and compare it to new methods.

The diligent but naive user might attempt to preserve this particular execution by saving the input file `data` in a digital repository, then making a note of the unique identifier of the data, and the exact version of `mysim` used in the published paper. In principle, the reader of the paper must simply install the given version of the software, download the data, and will quickly be able to verify, extend and compare with the published work. Unfortunately, this procedure is insufficient. The main problem is that what is visible to the user is only the tip of the iceberg in terms of what is necessary to actually execute the program.

Figure 1 gives a better sense of what may be involved in preserving such an execution. The binary executable `mysim` obviously depends upon `data` as an input file, but perhaps it also reads a file of calibration data `calib` in the current directory which is not mentioned on the command line, but hard-coded into the program. Further, the executable program itself does not stand alone, but depends upon a specialized library `libsim` that the user had to install onto the machine at some point. What's worse, the program itself makes a network connection at runtime in order to download some critical configuration data `config` from a public web server.

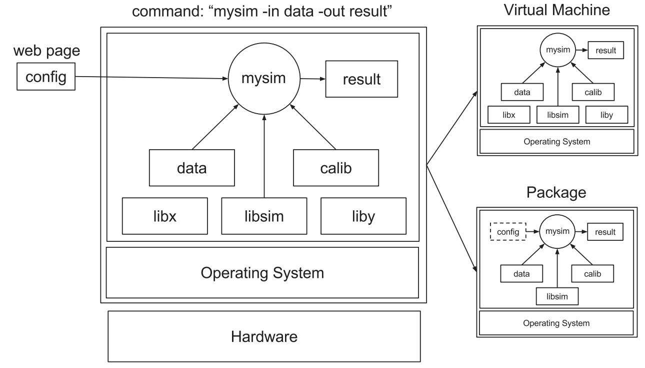


Figure 1: Preserving Implicit Dependencies
Even the simplest of programs has both explicit and implicit dependencies. Saving the program in a virtual machine tends to capture unnecessary items, while automatic packaging can identify exactly what objects are needed at runtime.

Even that is not the whole story. Even the simplest software depends on a complex stack of objects present on the local machine, including libraries, scripts, configuration files, and the operating system kernel. Together, these comprise what we call the **environment** of the program. While these components are of course required for the software to run, they are not the primary interest of the user, who cares first and foremost about the simulation and the data. In principle, the simulation should run correctly and yield the same results when run on a different (but compatible) operating system. In practice, it might not, and so preserving the environment is necessary for long-term viability.

Effective preservation requires that there exist some form of hardware capable of running the operating system and software. This could be physical preservation of a hardware artifact, or a compatible virtual implementation. Hardware preservation makes it easy to reproduce an application, but is not efficient due to the cost and space overhead for maintaining old hardware. At some point, the preserved hardware may become completely unusable due to humidity or the lifetime of components like disks. A compatible virtual implementation of old hardware, such as Olive [26], recreates the original execution environment on the future platforms through virtualization techniques [23].

An additional complication is that the different layers of the system may be provisioned by different parties. In the case of a personal laptop, the same person purchased the hardware, installed the operating system, and ran the software. But in a complex university computing environment, the hardware procurement, operating system installation, and software deployment may all be accomplished by multiple teams of people. By the time the end user gets involved, they may have no idea what the underlying environment

[1] http://www.daspos.org

actually contains!

The essence of the software preservation problem is that it is extremely difficult for the end user to understand the set of objects upon which an execution depends. The visible user interface suggests that the only required components are `mysim` and `data`, but the reality is that the program cannot run without a complex and interdependent set of invisible objects. Unless some additional specification or restrictions are put in place, any file on the local filesystem or any service available on the Internet could be a potential dependency of the execution. A preservation solution must either automatically capture what is unseen ("capturing the mess") or structure the user's interactions to make all dependencies explicit ("encourage cleanliness").

3. PRESERVATION OBJECTIVES

(Many terms are used in the field of digital preservation, including reproducibility, re-use, re-creation, re-purposing, and more, each with slight variations in meaning. To avoid confusion, we limit our terms to **preservation** to denote digital preservation whose purpose includes **verification** of previous results and **extension** to new results.)

Before posing solutions, it is useful to consider how the preserved software execution may be re-used in the future. It is commonly stated that researchers wish to precisely reproduce other's work so as to verify the truth of published claims [20]. In discussing the matter with a variety of researchers, we have found little appetite for attempting to prove or disprove other's work in this way. Rather, there are a wide variety of other motivations for precise reproducibility, most of them in the realm of reducing the amount of labor required to continue forward from a previous result. Examples include:

- **Identical Verification.** The same software executes on the same input data in the same environment and is repeated to verify that it produces the same result. This is done to evaluate the soundness of the reproduction system itself before moving on to other matters.
- **New Environment Verification.** The same software executes on the same input data in a **new** environment to verify that it produces the same result. This approach is taken to evaluate the soundness of new libraries, operating systems, hardware, and other parts of the environment as they evolve independently of the scientific objectives.
- **New Software Verification.** A new version of the same software executes on the same input data in the same environment, so as to verify that an improved implementation of the same algorithm yields the same results as the old.
- **Extension to New Data.** The same software executes on **new** data in the same environment. This allows previously published techniques to be extended to new data sets with confidence that new results are not affected by changes to the software or environment.
- **Extension to New Software.** Completely different software executes on the same data in the same environment. This allows for the direct comparison of different or competing algorithms on identical data, with confidence that the new publication has accurately reproduced the competing result.

Each of these use cases (except the first) requires a clear separation between the scientific software, the scientific data, and the computing environment, so that each can be evolved independently without accidentally modifying the other.

4. PRESERVING THE MESS

We first consider "preserving the mess" approaches, in which we attempt to capture exactly what the user attempted, without interfering in the setup of their work.

4.1 Virtual Machine Technology

A commonly-proposed solution is that software executions should be preserved by placing the software, data, and environment within a single virtual machine, then preserving the machine image in a repository either before or after the execution. This is a relatively easy technique for the user to apply, as long as the boundaries of the application correspond to the boundaries of a single machine and filesystem. If the application only depends upon objects in the local filesystem, each will be preserved at the bit-level in a precise way. Virtual machine preservation is effective and is already being used today at a small scale to capture individual complex systems [12].

However, when we consider preserving a large number of results that may evolve over time, virtual machine preservation has some significant limitations:

- **Imprecise Capture.** A virtual machine image will almost certainly contain items that are irrelevant to the task at hand. For example, a standard operating system contains a wide variety of software to handle many different user needs, most of which are not used by a given execution. Even worse, if the user preserves the image of their personal laptop, it could be all to easy to accidentally preserve personal data or legally sensitive information. On the other hand, the machine image by itself may fail to capture external dependencies (such as the `config` file on the web server) that are not strictly within the image, causing re-use to fail if the external dependency is not present.
- **Rigid Composition.** A virtual machine image intermixes the various components of the system in ways that are difficult to undo automatically. Absent some additional specification, there is no automatic way to distinguish the inputs to the simulation from the files comprising the application or the operating system. Manual effort to browse the image is the only way by which items can be extracted from the machine image.
- **Inefficient Storage.** It is rare for a single software execution to have scientific validity on its own. Rather, it is common for a researcher to run thousands to millions of instances of an application on a high-throughput computing system, each one using a slightly different input file or parameters. If we attempt to preserve each

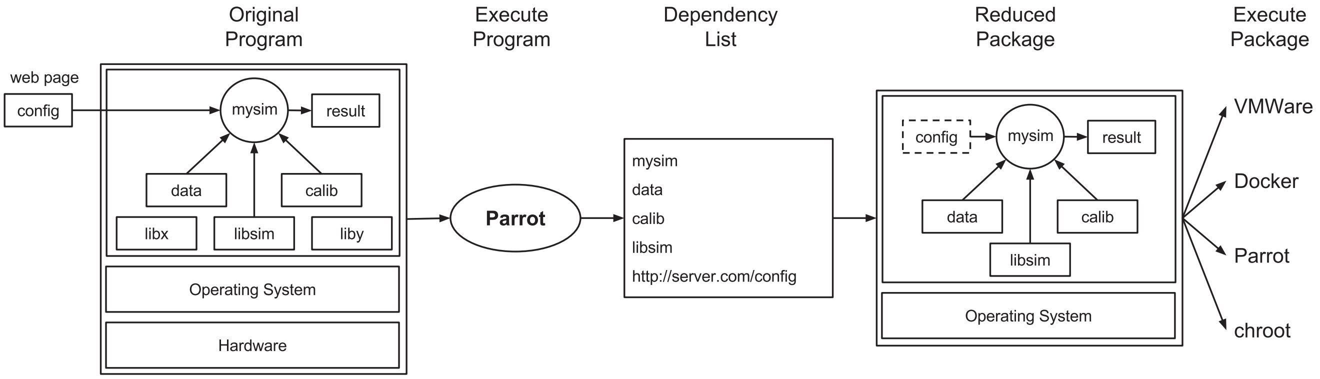


Figure 2: Packaging an Application with Parrot
Parrot can be used to trace the files and network objects used by a conventional program, and produce a listing of the items on which it depends. This listing is used to create a reduced package that can be re-executed by multiple technologies.

instance of the application in its own virtual machine, an enormous amount of storage will be consumed by duplicating the software, environment, and other components that are common to each instance.

- **Huge Image Size.** Data-intensive applications may have enormous input data sizes, measured in terabytes to petabytes. At this scale, the input data may be too large to store on a local disk, or to fit within a single virtual machine image. Large data sources are typically handled by purpose-built archives, and it is more effective for the virtual machine to refer to the archive than to duplicate its functionality.
- **Inefficient Execution.** There is sometimes an assumption that with a virtual machine "performance is of secondary importance" [14]. While this can be appropriate in some cases, users tend to stretch the limits of the available hardware to perform increasingly complex analyses. If a preservation method causes too much of a performance hit users will be unlikely to consider it until after getting their work done, if ever.

We conclude that the simple method of capturing a virtual machine image – while it may be useful – will not be an effective long-term strategy for preserving scientific software and data in a way that facilitates verification and extension.

4.2 Container Technology

Container technology is a growing alternative to hardware virtualization. Multiple containers can execute simultaneously on a single operating system kernel, and have lower execution overhead because they run directly on the CPU without translation or interception. Linux Containers (LXC), Rocket [1] and Docker [18] are examples of current systems that use this technology.

The stored image of a container is merely a stored filesystem tree. it may be stored as a disk image for efficiency, but can easily be exported in a portable, shareable format such as `tar` or `zip`. A container image can be a large, completely functional operating system with multiple applications, but users of container systems are encouraged to make small, minimal container images that support a single application at a time. However, the user must have enough understanding of the underlying application in order to construct the minimal image.

Although containers differ from virtual machines in the technology of execution, the container images themselves have the same problems as saving a filesystem image in a virtual machine, specifically imprecise capture, rigid composition, and inefficient storage. To use either technology effectively, the user needs additional help to identify dependencies.

4.3 Package Reduction with Parrot

Tracing techniques can be used to determine the minimal set of objects needed to support an application, and then use that information to construct an appropriate package of actual dependencies in either a virtual machine image or container image. A monitor process can run alongside an executing instance of an application, observe its interactions with the environment, and then save only those elements of the environment into a new package. A variety of technologies can then be used to re-execute the software.

Parrot [16] is an example of this technique, which is also employed by CDE [10], and PTU [21]. Parrot was originally designed to be a remote filesystem access tool which connects conventional applications to remote I/O systems such as HTTP and FTP. It works by trapping system calls through the `ptrace` interface and replacing selected operations with remote accesses. Through this technique, Parrot is able to modify the filesystem namespace in arbitrary ways according to user needs. Parrot is particularly used in the high energy physics community to provide remote access to application software via the CVMFS [4] file system.

To support package creation, we made small modifications to Parrot to record the logical name of every file accessed by an application into an external dependency list. After execution is complete, a second tool is used to copy all of the named dependencies into a package. In addition, Parrot tracks the network operations of an application and the data passing through them. It records the address, port number, and protocol of each connection. In addition, it examines each connection for known protocol signatures and can determine the protocol-level endpoint of the connection. For example, if the application connects to a webserver, Parrot can record not only the address of the webserver, but also the URL which the application accessed for common protocols such as HTTP, SVN, and GIT. (Parrot is limited in that it cannot inspect encrypted data, beyond indicating that a TLS/SSL connection was made.)

Figure 2 gives an overview of how a package is made. First, the user executes the program in the normal way, using Parrot. The application runs to completion while Parrot collects the files and URLs accessed into a **dependency list**. All the accessed files are copied into a package so that the file system structure (relevant paths between files, and symbolic links) is kept within the package. The package is a simple `tar` archive that can be recorded in any digital repository and then re-executed by a variety of techniques. For example, the package can be converted into a virtual machine image and executed by VMWare [23], or it can be converted into a container image and executed by Docker [18]. Parrot itself can also be used to re-execute the package by mounting the package directory as the application's root directory.

The reduced package is certainly smaller than the entire virtual machine image, but can still be astonishingly complicated. As shown in Figure 2, all the file dependencies, including files from the root filesystem like `/bin` and files from the network filesystems like AFS and CVMFS, are denoted as file paths within the dependency list. The distinction between input data and software is lost, which makes extensions based on a preserved package difficult. In addition, common library dependencies will be wrapped into different packages multiple times, which increases the storage overhead of the remote archive. In an earlier work, we used Parrot to preserve a simple high energy physics application called `TauRoast`. The reduced package contained 22,068 files and directories totaling 21 GB of data and software drawn from 8 different filesystems. Virtually all of this detail was unknown to the invoking user.

Based on this experience, we believe that these approaches are ultimately limited. While "preserving the mess" is better than not preserving at all, the resulting packages are extraordinarily complicated, and provide the end user with little traction for understanding the behavior well enough to extend the software. Preserving the mess is inherently retrospective – it involves observing an execution after it executes to infer what resources were consumed. A more structured approach is needed for extending the original work.

5. ENCOURAGING CLEANLINESS

In contrast to preserving the mess, "encouraging cleanliness" is a forward-looking approach. Cleanliness is accomplished by encouraging everyone to name and preserve objects **before** they are used, then to combine the objects at runtime in a way that clearly distinguishes the reusable layers of the application. To support cleanliness, an archive is needed to maintain the OS images, software, and data for each software execution. A specification should be created to describe the execution environment for each execution with the help of the system administrator and the original author.

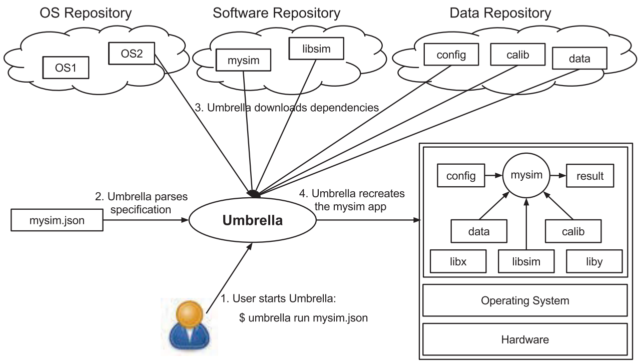


Figure 3: Overview of Umbrella
Umbrella is used to execute a specification of an application which describes precisely how the operating system, software, and data are combined at runtime.

Here, we demonstrate two approaches to cleanliness: Umbrella preserves the execution of a single software execution by precisely naming the hardware, operating system, software, and data necessary to carry it out. PRUNE preserves the execution of a workflow of software executions by preserving multiple software executions independently, then using Umbrella to execute each one precisely.

5.1 Precise Execution with Umbrella

Umbrella [17] is designed to enable the precise construction of an execution environment for software. Figure 3 gives an overview of the system. The user gives a declarative specification of the desired execution environment, encompassing the hardware, kernel, OS, software, data, and environment variables, without being tied down to a single virtualization technology. Umbrella considers each of the elements of the specification, downloads the files needed, constructs the complete environment by combining the components, then runs the program.

Figure 4 gives a possible Umbrella specification for our example program. The `hardware` section indicates the required CPU architecture, the CPU model, the CPU flags, the number of cores, and the amount of memory, disk and other hardware requirements. The `kernel` section defines the type and version of the operating system kernel, which may be a single value or a range. The `os` section provides the name and version information of the operating system, which includes the system software in the root filesystem, apart from the kernel. The `software` section provides the software name, version, and platform of each required software package. The `data` section indicates the necessary data dependencies, and their mount points. The `environ` section sets the environment variables for an application. For each category, a variety of methods of naming the object are available, ranging from **unique identifiers** (`id="e5f3cd"`) to **abstract attribute values** (`version="6.5"`), depending on what is most appropriate for the user. The user may select whichever method best meets their needs. We discuss tradeoffs in naming schemes at length below.

Note that Umbrella requires the user to be explicit about external dependencies. As our example shows, the exter-

```
"hardware": {
    "platform": "x86_64",
    "cpu cores": "1",
    "memory": "1 GB",
    "disk": "4 GB"
},
"kernel": {
    "type": "Linux",
    "release": "2.6.32"
},
"os": {
    "name": "GreenSock",
    "version": "8.3"
},
"software": {
    "mysim": {
        "id": "f6e17cc80...",
        "mountpoint": "/software/mysim",
        "version" : "3.2"
    }
},
"data": {
    "config": {
        "url" : "http://server.com/config",
        "mountpoint": "/etc/mysim/config"
    },
    "data": {
        "id": "cb9878132...",
        "mountpoint": "/home/test_user/mysim/data"
        }
    },
},
"environ": {
    "HOME": "/home/test_user",
    "PATH": "/usr/bin:/software/mysim/bin"
},
"command": "mysim -in data -out result"
```

Figure 4: Example Umbrella Specification
This example of an Umbrella specification indicates exactly how the components of `mysim` *come together to form a complete execution.*

nal web page containing `config` is explicitly mentioned, so that Umbrella itself will download the data and provide it to the application. The user may make a value judgement about the long-term availability of the external dependency. To avoid data loss, `config` should be archived into the data repository, together with its metadata including its checksum, size, authorship, access permission and usage. The specification of `mysim` should include `config` as one of its data dependencies through its unique identifier or attribute list. Similarly, the stability and persistency of all the third-party dependencies should be evaluated, and the unstable ones should be ingested into the archive if access permission is allowed. Once all items are archived, then the specification itself is a (compact) archivable object that completely describes the execution.

The Umbrella specification is deliberately silent about the specific *mechanism* by which the program will be re-executed. This gives the implementation freedom to make use of new technologies as they are developed, or to harness whatever resources are available at the moment of execution. For example, if Umbrella is invoked on a machine that already has the desired operating system on compatible hardware, then it can simply run the software directly. If the hardware is compatible but the OS is not, then Umbrella can attempt to use a container to deploy the desired OS. If not even the hardware is compatible, then Umbrella can instantiate a virtual machine or contact a commercial cloud service to create the desired environment.

The specification is inherently efficient in both use of storage space, and in construction of the desired environment. Each of the components in the specification is assumed to be preserved in an external digital repository, then downloaded and cached at the execution site as needed. Obviously, if multiple executions use the same operating system or the same dataset, it is only necessary to keep one copy in the archive and share it at runtime among multiple executions.

Previous approaches to the provisioning of virtual machines, such as V-MCS [28], FutureGrid [30], Grid'5000 [6], and VMPlants [13], achieve various environments by applying executable scripts to base virtual machines. While effective, this can be quite slow while data is copied or updated in place. In contrast, Umbrella *mounts* each object in the filesystem namespace, so that at runtime, the collection of objects is effectively instantaneous.

5.2 Preserving Workflows with PRUNE

While Umbrella describes how to precisely perform a *single* software execution, PRUNE describes how to connect *multiple* executions together, such that entire workflows can be preserved, verified and extended. The key idea of PRUNE is to represent every invocation of a program as the evaluation of a function on immutable digital objects. In PRUNE, our example program would be invoked as a function call:

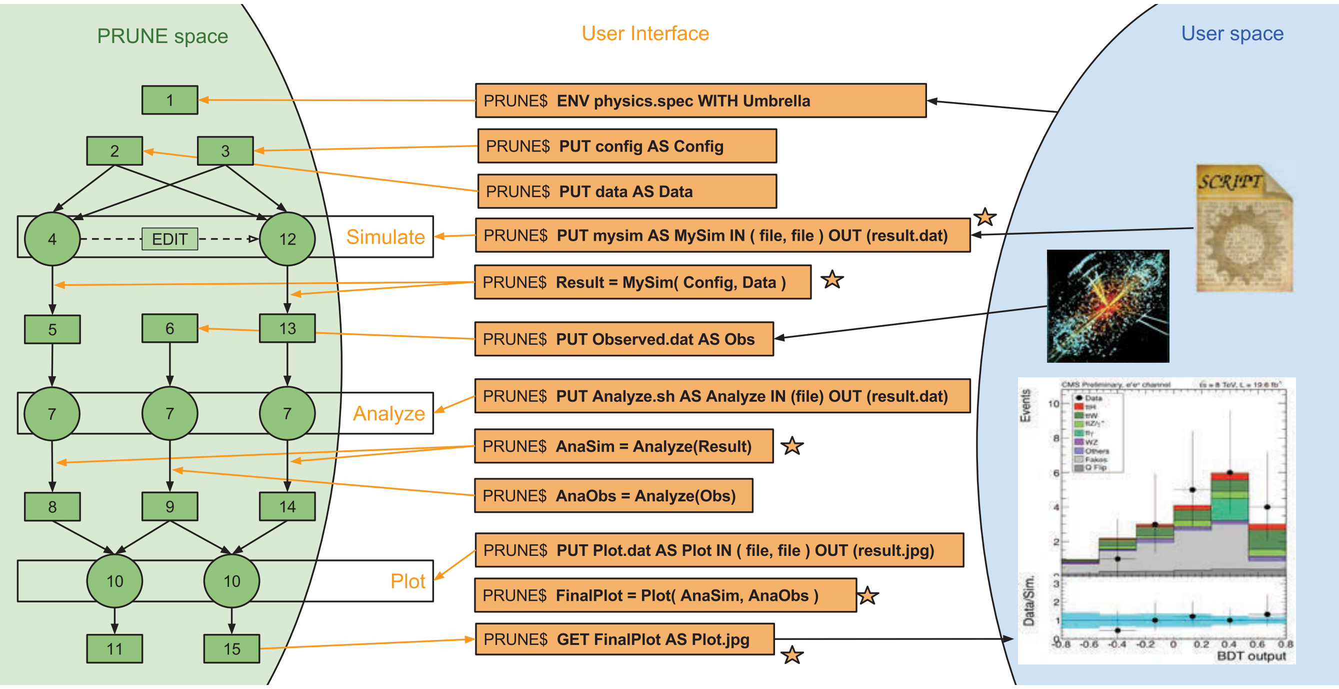
```
Result = MySim( Config, Data )
```

In this example, `Config` and `Data` refer to data items stored in "PRUNE-space", a local repository of immutable objects. `mysim` consists of an Umbrella specification of how to execute the program in a precise environment, while `Result` refers to the output file, which is moved into PRUNE-space when the program completes.

Over time, as the user runs a large number of programs, they conceptually build up a large graph of objects, each related to each other by function invocations. If an object was created by a chained series of function calls, PRUNE retains enough information to accurately describe the steps necessary to create that object from beginning to end. The user who wishes to publish a paper depending upon a result can ask PRUNE to produce a package containing every dependency needed for that result, which can then be archived along with the scientific publication.

Of course, accumulating those objects over time will exhaust disk storage, or the user's budget for cloud storage. To this end, PRUNE can safely delete the binary form of any object in PRUNE-space, because it retains enough information to re-create it, should the user require that it be produced. In this way, storage costs can be traded for computation costs as needed.

PRUNE gives all objects a uuid, but managing a large number uuids manually would quickly become cumbersome. So each repository in PRUNE has it's own namespace on top of the uuids such that a name points to a uuid, and both the name and uuid are preserved. A collaborating reposi-



```
PRUNE$  ENV physics.spec WITH Umbrella
PRUNE$  PUT config AS Config
PRUNE$  PUT data AS Data
PRUNE$  PUT mysim AS MySim IN ( file, file ) OUT (result.dat)
PRUNE$  Result = MySim( Config, Data )
PRUNE$  PUT Observed.dat AS Obs
PRUNE$  PUT Analyze.sh AS Analyze IN (file) OUT (result.dat)
PRUNE$  AnaSim = Analyze(Result)
PRUNE$  AnaObs = Analyze(Obs)
PRUNE$  PUT Plot.dat AS Plot IN ( file, file ) OUT (result.jpg)
PRUNE$  FinalPlot = Plot( AnaSim, AnaObs )
PRUNE$  GET FinalPlot AS Plot.jpg
```

Figure 5: PRUNE overview

Prune represents each invocation of a program as a function call on immutable archived objects. As the user invokes more and more functions, a tree of archived objects accumulates. Each execution is made precisely reproducible via Umbrella.

tory can choose to use the same name, or not, but uuids are immutable across repositories.

In PRUNE, a distinction is made between *operations* which are specified programmatically, and *edits* that are transformations performed manually and might not even have a detailed description. Precise reproducibility is possible in all cases, but including an edit in the workflow could leave a gap in the provenance and make it difficult for a collaborator to reproduce an edit on a file if the original file has changed.

An edit which does not leave a gap in the provenance is shown in Figure 5 from object 4 to 12. This edit allows to user to easily manage minor evolutionary changes to the workflow without harming the ability to preserve the workflow for collaborators. Notice that tracing 11 or 15 back to their original source files does not require passing through the edit.

6. PRESERVATION CONSIDERATIONS

Within the overall strategies outlined above, a number of tradeoffs become apparent between user effort, preservation cost and complexity, and the generality of the artifacts for re-use over the long term. Here we give some overview of these tradeoffs and suggest when one approach or the other (or both) may be appropriate.

6.1 Source vs. Binary Code

Should we preserve the source form or the binary form of compiled programs?

By design, source code in a high level language such as C is designed for human consumption and is the preferred form for understanding and modifying the program. The binary form produced by the compiler can be directly executed but is of little use for analysis and may not function in even a slightly different environment. Source code is obviously the preferred form in which software itself achieves longevity as an independent entity.

However, if our goal is to preserve an *instance of executed software*, the answer is not so clear. If only the source code is preserved, then re-use requires that a suitable compiler, linker, and other supporting tools be present at re-use time. Languages are not always forward compatible, which requires us to preserve the actual compiler used in addition to the software. Moreover, the source code for the compiler also needs to be preserved, and so on. As the dependency chain increases, the cost of re-execution increases, as well as the risk of failing to build a usable binary.

Thus, the comprehensive approach is to preserve both the source code and the resulting binary, so that they are mutually verifiable. If the surrounding environment is faithfully preserved, then the binary will remain usable. If it is desired to rebuild the software from source, the correctness of the rebuild can be confirmed by comparing its outputs against those generated by the preserved binary. Similar arguments apply to any complex object constructed from text instructions, such as an RPM package from a rpmbuild script or a Docker image from a Dockerfile.

6.2 Manual vs. Automatic Preservation

Should preservation be performed automatically, or only at the user's request?

Automatic preservation does not need lots of involvement from the user, but can be very messy. Since everything is recorded, irrelevant operations become part of the preserved data, which are difficult to distinguish from the relevant operations. The irrelevant operations may include listing files in a directory or iterations that failed to produce the desired results and had to be modified and re-run. Operations at lower levels may be difficult to decipher for even the original researcher. It is nearly impossible for another researcher to use this type of data to do extension work.

Automatic preservation may also cause privacy issues for the researcher. A preservation tool which is allowed to track the software execution may preserve the researcher's ssh private key and private key file including his Amazon EC2 key pair. Distributing the preserved execution may leak the private information of the researcher to unintended targets.

At the other extreme, manual preservation places the entire burden of reproducibility on the researcher. The researcher might create a script that includes the final list of operations they used to produce the experimental results. Or the researcher might create documentations explaining each step with details about why certain decisions were made. Or the user might include very little information about how the results were obtained, making the preservation ineffective.

6.3 Pre vs. Post Preservation

Should the burden of preservation come before or after the user's work is completed?

Most researchers choose to wait until they have their full results before preserving their methods. Unfortunately, by the time the results are available, other factors come into play which make this approach unlikely to succeed. There is little motivation to put in the extra effort to identify how results were obtained since it does not appear to be a factor in whether or not a paper is accepted. The researcher typically gets busier as a deadline approaches and preservation has a low priority. In addition, sometimes important details are forgotten or only known by system administrators. In a collaborative effort, students who were involved may have moved on by that time. Also, the original execution environment may have changed, and the original operations no longer work.

An alternative approach is to require the researcher to preserve every step along the way before it is even executed. This would require more work for the researcher up-front, but has a much higher likelihood of resulting in a preserved software execution. However, this approach would also include failed attempts or extraneous commands that occur as the research evolves. This extra data puts additional load on resources used to capture and store the information.

A middle ground can be found by provisionally preserving everything in a local repository, and then enabling the user to identify (at a later time) what objects should be retained permanently. This requires more effort from the researcher both before and after the execution, but provides a clean description of how the research can be verified, extended and compared. Furthermore, the extra burden on the user might be offset by providing some additional tools with convenience features.

6.4 Unique Identifiers vs. Attributes

How should preserved software components be identified?

Each component of a software execution should allow the user to refer and verify its integrity. There are two broad approaches for this: **unique identifiers** or **attribute descriptions**. The following are examples of each type.

```
Unique Identifiers:
    doi = "10.7274/R0C24TCG";
    checksum = "f6e17cc80...";
    url = "http://server.com/config";

Attribute Descriptions:
    name = "mysim";
    version = "3.2";
    architecture = "x86_64";
```

Unique identifiers provide an unmistakable reference to a single binary object. A Digital Object Identifier (DOI) [19] is an example that names a publisher, then an object, which can be resolved by the Handle System [27] to a current location of the desired object, in the form of a Uniform Resource Locator (URL). DOIs are widely used by digital libraries to identify published documents, and to a lesser extent, other kinds of digital objects. However, the DOI system recommends, but does not force, the objects referred by a DOI name to be persistent or immutable. For example, the DOI name `10.1000/182` always refers to the *latest* version of the DOI handbook, which is the primary source of information about the DOI system. [2]

Attribute descriptions describe essential properties of the software but do not necessarily uniquely identify a image. A suitable set of attribute-value pairs can be used to search a known repository for corresponding images that satisfy the given requirements, and are likely to resolve to a small set of compatible objects.

It seems that despite all of the advances that an internet connected research community provides, the question of archiving identification information is an issue yet to be resolved [29]. As long as the unique identifiers are kept immutable, they could be used as persistent identifiers in a global system of identifiers [2]. And a similar system focused on the evolving translations could provide both user friendliness and reproducibility.

[2] As an aside, the DOI infrastructure is almost, but not quite, suitable for naming software components. The main problem is that DOIs generally resolve to a web page that describes the "concept" of the object for a human reader, rather than resolving to the binary object itself. What is needed is a unique name that resolves directly to the object, or a convention for resolving the object from the concept page itself.

7. RELATED WORK

The problem of software verification is not new. As far back as 1984 [5], efforts were made to encourage verification that design specifications matched the actual behavior of software. However, the paper demonstrated that when the guidelines are followed, deviations from the specifications can be detected earlier, saving time in the overall software development. This tight coupling of specification and implementation has benefits for preservation also.

Dendro [8] has the user do some work as early as possible in order to preserve the provenance. For example, rather than just providing a link to a website, a triple is used to describe that the URL is the creator's web page. Using this ontology based data model rather than relying on a relational database, the preservation becomes self-documenting. A system called dataref versuchung [9] also requires some upfront work by the researcher. But once the researcher has properly created a figure for a LaTeX document, the system automatically includes a datagraphy that includes information about how the figures were created.

Unlike the above approaches which require significant user intervention, the PERICLES Extraction Tool [7] is initiated at runtime and attempts to automatically detect all implicit dependencies on the system environment and convert them to explicit dependencies. It also attempts to rank significant events to make the result more organized. However, it is still possible that this tool may miss implicit dependencies introduced by the extension work.

Matthews [15] proposes a conceptual framework for software preservation which includes a performance model of software and its input data, a model of software components, and the categories of preservation properties of software such as functionality, composition, provenance, ownership, execution environment and so on. Hong [11] proposes a benefits framework for software preservation which enumerates different purposes of software preservation and its benefits, analyzes the pros and cons of integrating software preservation measures into software development processes and preserving legacy software separately, and provides different options for software preservation. In contrast to this research on software preservation, our work considers the preservation of scientific software executions systematically, which includes data, software, and execution environments.

8. OPEN PROBLEMS

In this paper, we have outlined what we see as the most pressing problems of digital preservation and outlined broad strategies for solving them. There remain many hard problems to consider:

Preservation of Distributed Applications. Compared with single-machine applications, the preservation of distributed applications is more challenging due to the following facts: First, a distributed system is often composed of multiple computer nodes, each has its own software stack. Second, the distributed model and the network configuration must be maintained to reconstruct the distributed systems. Third, some distributed systems like HTCondor are dynamic, in that nodes can join and leave the Condor pool at any time. Should we preserve distributed applications including software and hardware completely? Should we just preserve the detailed configuration of distributed applications? Should we only preserve the working principle of distributed applications?

Preservation Granularity. To preserve applications, there are three different kinds of packages involved in an archive: Submission Information Packages (SIPs), Archive Information Packages (AIPs) and Dissemination Information Packages (DIPs) [31]. The granularity for these packages may be different. For example, an archive may choose to split submission packages into smaller pieces to fit its underlying storage architecture. The choice of granularity depends on the overhead of metadata management, storage overhead, the time overhead of submission, storage and reconstruction, and user-friendliness.

Preserving Preservation Tools. Emulation, as an important preservation approach, emulates the original execution environment of an application to allow the application to execute without modification. Compared with migration, where every preserved application needs to be somehow modified to fit the new environment, emulation keeps all the applications unchanged, and emulates the previous execution environment [14, 24].

Commercial Software and Sensitive Data. Preserving both the source code and binary code can help the new users extend the work easily. However, sometimes it is difficult to get the source code of software, especially commercial software. The preservation policy for this type of software must take copying and distribution conditions into consideration. On the other hand, trapping system calls may expose some sensitive data, which should require special access permissions. Before wrapping all these data into a reduced package, the sensitivity of the preserved data should be considered.

Acknowledgements

This work was supported in part by National Science Foundation grant PHY-1247316, Data and Software Preservation for Open Science (DASPOS).

9. REFERENCES

[1] CoreOS is building a container runtime, Rocket. `https://coreos.com/blog/rocket/`, 2015.

[2] B. Bazzanella. A persistent identifier e-infrastructure. *IPRES 2014 proceedings*, page 118, 2014.

[3] C. G. Begley and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012.

[4] J. Blomer, P. Buncic, and T. Fuhrmann. CernVM-FS: delivering scientific software to globally distributed computing resources. In *Proceedings of the first international workshop on Network-aware data management*, pages 49–56. ACM, 2011.

[5] B. W. Boehm. Verifying and validating software requirements and design specifications. In *IEEE software*. Citeseer, 1984.

[6] R. Bolze, F. Cappello, E. Caron, M. Daydé, F. Desprez, E. Jeannot, Y. Jégou, S. Lanteri, J. Leduc, N. Melab, et al. Grid'5000: a large scale and highly reconfigurable experimental grid testbed.

International Journal of High Performance Computing Applications, 20(4):481–494, 2006.

[7] 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. *IPRES 2014 proceedings*, page 249, 2014.

[8] J. R. da Silva, J. A. Castro, C. Ribeiro, and J. C. Lopes. The dendro research data management platform. *IPRES 2014 proceedings*, page 189, 2014.

[9] C. Dietrich and D. Lohmann. The dataref versuchung: Saving time through better internal repeatability. *ACM SIGOPS Operating Systems Review*, 49(1):51–60, 2015.

[10] P. J. Guo and D. R. Engler. CDE: Using System Call Interposition to Automatically Create Portable Software Packages. In *USENIX Annual Technical Conference*, 2011.

[11] N. C. Hong, S. Crouch, S. Hettrick, T. Parkinson, and M. Shreeve. Software preservation benefits framework. *Software Sustainability Institute Technical Report*, 2010.

[12] G. Juve, E. Deelman, K. Vahi, G. Mehta, B. Berriman, B. P. Berman, and P. Maechling. Scientific workflow applications on Amazon EC2. In *E-Science Workshops, 2009 5th IEEE International Conference on*, pages 59–66. IEEE, 2009.

[13] I. Krsul, A. Ganguly, J. Zhang, J. A. Fortes, and R. J. Figueiredo. Vmplants: Providing and managing virtual machine execution environments for grid computing. In *Supercomputing, 2004. Proceedings of the ACM/IEEE SC2004 Conference*, pages 7–7. IEEE, 2004.

[14] R. A. Lorie and R. J. van Diessen. UVC: A universal virtual computer for long-term preservation of digital information. *IBM Res. rep. RJ*, 10338, 2005.

[15] B. Matthews, A. Shaon, J. Bicarregui, and C. Jones. A framework for software preservation. *International Journal of Digital Curation*, 5(1):91–105, 2010.

[16] H. Meng, R. Kommineni, Q. Pham, R. Gardner, T. Malik, and D. Thain. An invariant framework for conducting reproducible computational science. *Journal of Computational Science*, 9:137–142, 2015.

[17] H. Meng and D. Thain. Umbrella: A portable environment creator for reproducible computing on clusters, clouds, and grids. In *Proceedings of the 8th International Workshop on Virtualization Technologies in Distributed Computing*, VTDC '15. ACM, 2015.

[18] D. Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux Journal*, 2014(239):2, 2014.

[19] N. Paskin. Digital object identifier (DOI) system. *Encyclopedia of library and information sciences*, 3:1586–1592, 2008.

[20] R. D. Peng. Reproducible research in computational science. *Science (New York, Ny)*, 334(6060):1226, 2011.

[21] Q. Pham, T. Malik, and I. T. Foster. Using provenance for repeatability. In *USENIX NSDI Workshop on Theory and Practice of Provenance (TaPP)*, 2013.

[22] F. Prinz, T. Schlange, and K. Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? *Nature reviews Drug discovery*, 10(9):712–712, 2011.

[23] M. Rosenblum. Vmware?s virtual platform? In *Proceedings of hot chips*, volume 1999, pages 185–196, 1999.

[24] J. Rothenberg. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation. A Report to the Council on Library and Information Resources.* ERIC, 1999.

[25] B. Sierman. The scape policy framework, maturity levels and the need for realistic preservation policies. *IPRES 2014 proceedings*, page 259, 2014.

[26] G. St Clair and D. Ryan. Olive: A digital archive for executable content. *Coalition for Networked Information*, 2011.

[27] S. Sun, L. Lannom, and B. Boesch. Handle system overview. Technical report, RFC 3650, November, 2003.

[28] X.-H. Sun, C. Du, H. Zou, Y. Chen, and P. Shukla. V-mcs: A configuration system for virtual machines. In *Cluster Computing and Workshops, 2009. CLUSTER'09. IEEE International Conference on*, pages 1–7. IEEE, 2009.

[29] H. Van de Sompel and A. Treloar. A perspective on archiving the scholarly web. *IPRES 2014 proceedings*, page 194, 2014.

[30] G. Von Laszewski, G. C. Fox, F. Wang, A. J. Younge, A. Kulshrestha, G. G. Pike, W. Smith, J. Voeckler, R. J. Figueiredo, J. Fortes, et al. Design of the futuregrid experiment management framework. In *Gateway computing environments workshop (GCE)*, pages 1–10, 2010.

[31] E. Zierau and N. Y. McGovern. Supporting the analysis and audit of collaborative oais?s using an outer oais-inner oais (oo-io) model. *Preservation (DDP)*, 9:5.

Towards a Common Approach for Access to Digital Archival Records in Europe

Alex Thirifays
Danish National Archives
Rigsdagsgården 9
1218 København K
+45 41 71 72 47
alt@sa.dk

Kathrine Hougaard Edsen Johansen
Danish National Archives
Rigsdagsgården 9
1218 København K
+45 41 71 72 20
khej@sa.dk

ABSTRACT

This paper describes how the E-ARK project (European Archival Records and Knowledge Preservation) aims to develop an overarching methodology for curating digital assets. This methodology must address business needs and operational issues, proposing a technical wall-to-wall reference implementation for the core OAIS flow – Ingest, Archival Storage and Access.

The focal point of the article is the Access part of the OAIS flow. The paper first lays out the access vision of the E-ARK project, and secondly describes the method employed to enable information processing and to pin-point the functional and non-functional requirements. These requirements will allow the E-ARK project to create a standardized format for the Dissemination Information Package (DIP), and to develop the access tools that will process this format. The paper then proceeds to describe the DIP format before detailing what the access solution will look like, which tools will be developed and, not least, why the E-ARK Access system will be used and work.

General Terms

Frameworks for digital preservation, Preservation workflows, Infrastructure opportunities

Keywords

Access management, DIP format, E-ARK-project, digital archives.

1. INTRODUCTION

European National Archives have partnered up with vendors of digital preservation services, research institutions and interest groups to develop common ground for OAIS compliant [3] digital archiving. Common tools and common Information Packages (IP) are at the heart of this project. The differences in legislation and practices in place across Europe mean that workflows and processes cannot meaningfully be aligned. Using the same tools and IP formats will, however, lead to new possibilities for collaboration across boarders and national differences. Among the benefits of closer collaboration on core parts of digital archiving is the possibility of more cost effective use of resources that uniting of efforts can lead to.

An OAIS compliant end-to-end methodology for digital archiving with common formats for OAIS Information Packages and tools will be the outcome of the project. The objective is to provide a single, scalable, computational and robust approach capable of meeting the needs of diverse organisations, public and private, large and small, with capacity of supporting from simple to complex content data types. Common formats for SIPs, AIPs and DIPs are being developed along with supporting tools. The methodology will cover multiple content data types of which databases, ERMS data[1], geo-data, and multidimensional data sets (OLAP cubes[2]) for data mining are the most prominent.

The ambition is that the methodology becomes an international standard allowing archival institutions to hand-pick single components, e.g. access presentation tools, or plug-in the whole reference implementation and be ready to do OAIS compliant digital archiving.

The E-ARK initiative is a 3-year multinational research project that runs from 1st of February 2014 to 31st of January 2017. It is co-funded by the European Commission under its ICT Policy Support Programme (PSP) within its Competitiveness and Innovation Framework Programme (CIP). The consortium holds 16 partners from 11 countries and collaborates with archives in Sweden and Switzerland.

The project is divided into 8 Work Packages, and the Danish National Archives is leading the Work Package on Access.

More information about the project is available from the website at www.eark-project.eu.

2. A COMMOM APPROACH TO ACCESS

2.1 Filling a Gap

The ultimate goal of preservation actions must be to ensure that access and reuse is possible. However, experience related to providing access to born-digital archival material is still limited. A study carried out in 2014 early in the E-ARK project [1]

[1] Electronic Records Management System (ERMS)

[2] OLAP is an acronym for online analytical processing. It is a technique originating from Business Intelligence needs and is used for fast and in-depth analysis of large data sets that are logically arranged in so-called snowflake schemas.

confirmed this view. The study further revealed extensive gaps between the requirements for access and existing services for access [2].

Few archives provide access to databases, geo-data, ERMS data, OLAP cubes, and other complex born-digital materials. However, many archives have expressed the need to provide access to one or more of these complex content data types now or in the near future. When comparing user needs with existing access solutions the gaps are also obvious. Existing access solutions are not very user-friendly and generally only meet users' needs poorly. The most significant gaps in relation to users' needs are:

- Lack of functionalities in access tools to support intended use of data
- Lack of comprehensive and qualitative metadata in finding aids makes it difficult to find data of interest
- Lack of flexible and modern access services
- Lack of interoperability between access components

Throughout the project E-ARK will aim to develop solutions that will help bridge the identified gaps and improve possibilities for access to digital archival materials.

2.2 Powerful Access Tools

The vision in E-ARK is to create components for access and re-use that bridge the most significant gaps. Tools for providing access to born-digital archival records will be developed and a special focus will be on complex content data types. User-friendliness and flexibility of the tools are top priorities. The same goes for ensuring that tools are easy to use and will allow consumers to access and use material for their intended purposes. Behind the tools lies a robust, common DIP format that enables efficient access via the developed tools.

2.3 For the Benefit of Archives and Users

Archivists and end-users alike will benefit from a closer collaboration and use of common tools and formats across borders. For archives an obvious benefit is that much needed access tools for complex content data types will be available, but among the possible benefits are also efficiency and cost effective use of resources. The extensive focus on user-friendliness and the determination to create tools that meet the present needs for access services will bridge some of the largest gaps between existing solutions and users' needs. Users will benefit from modern digital archival access solutions developed with user requirements in mind. Further the implementation of a common DIP format across national archives introduces new perspectives for cross-border research in archival material.

3. METHOD

The approach adopted to identify requirements for a common DIP format and access tools was formalized in a requirements specification template designed to be used by all work packages dealing with the creation of IPs and development of tools (work package on Ingest, Archival Storage, Access and Services and Integration). The information processing approach was double and consisted in a bottom-up and a top-down identification of requirements.

3.1 Bottom-Up Approach

The bottom-up approach entailed detailed analyses of requirements from essentially three different sources: User; tools; and metadata standards.

3.1.1 User Needs for Digital Archival Access Solutions

First step was to study the existing landscape of access to digital archival materials, identify user needs for access services, and then find the gaps between existing solutions and user needs. This was done in spring 2014 where a survey and a series of qualitative follow-up interviews were carried out with a broad range of stakeholders [2].

The results of the gap analysis form an important foundation for the onward work and will be referenced continuously to ensure that what is developed is something which is in demand and, equally as important, that it will meet the quality goals of users.

3.1.2 Requirements for Pilot Sites and Access Tools in General

The E-ARK project plans to operationalize the tools at specific pilot sites prior to their release. It was therefore relevant for multiple purposes to examine the requirements of each pilot site and to include them directly in the requirements specification. The access pilot sites include: KEEP SOLUTIONS, the Estonian Business Archives, the National Archives of Estonia, the National Archives of Hungary, the National Archives of Slovenia, and most probably also the Danish National Archives. The foci of these pilot sites are different and the identified requirements therefore reflect this and also cover different services, legislations, and data types, such as business records and databases, CMIS access to single records, access to Moreq compliant Electronic Document and Records Management Systems (ERMS), and access to geo-data.

3.1.3 The Metadata Elements for the DIP

In order to identify which metadata elements were needed in the common DIP format, a number of existing metadata standards were chosen for examination: METS[3], PREMIS[4], apeEAC-CPF[5], GML (INSPIRE)[6], SIARD[7], BagIt[8], Dublin Core[9], and EAD[10].

To allow for comparison of the standards, it was agreed that the analysis of the standards should not be at the level of individual metadata elements, but rather in terms of semantic metadata categories. There are a number of possible ways in which metadata elements can be categorised and there is no single "right" way of doing this. The following categories were chosen to proceed with the analysis: Provenance, Context, Discovery, Relations, Rights, Reference Information, Preservation, Integrity, Storage, and Datatype.

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

[4] PREMIS (Preservation Metadata: Implementation Strategies) http://www.loc.gov/standards/premis/

[5] apeEAC-CPF (Encoded Archival Context for Corporate Bodies, Persons, and Families) http://eac.staatsbibliothek-berlin.de/

[6] GML(INSPIRE) http://wiki.osgeo.org/wiki/INSPIRE

[7] SIARD (Software Independent Archiving of Relational Databases)http://www.bar.admin.ch/dienstleistungen/00823/01911/index.html?lang=en

[8] BagIt https://tools.ietf.org/html/draft-kunze-bagit10

[9] Dublin Core http://dublincore.org/

[10] EAD (Encodes Archival Description) http://www.loc.gov/ead/

In order to produce a more detailed impression of the coverage of each standard, the number of metadata elements belonging to each of the categories was recorded. This quantification made it possible to highlight potentially interesting differences between the standards which could subsequently be further investigated by drilling down into the metadata elements of the standards themselves. In addition, investigators were asked to provide a brief description of how they counted the elements, the nature of the standard itself and any other comments relevant to comparing the standards.

These thorough and detailed examinations of user needs, requirements for tools, and metadata standards were the first part of the adopted method for specifying the E-ARK DIP format and the associated access tools.

The second part of the job was to employ a top-down approach to complete the identification of requirements.

3.2 Top-Down Approach

The top-down approach consisted in making a comprehensive breakdown of the activities that make up the access flow; identify use cases and attach acceptance criteria and constraints to these; as well as identifying functional requirements.

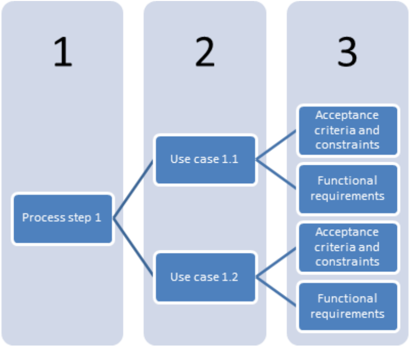


Figure 1. Decomposition of the Access flow

In the following sections, this top-down method will be described.

3.2.1 High-level Illustration and Process Step Description

Creating high-level illustrations and descriptions of the generic process steps for the whole access flow has contributed to reaching a common understanding between users (archivists) and developers. It has also defined the scope of the access activities that need to be underpinned by tools developed in E-ARK. Furthermore this work has been used to create a first platform for discussion, enabling the identification use cases.

Both high-level illustrations and descriptions are based on the E-ARK General Model, which sets up a common conceptual framework for the entirety of digital archival activities.

The overall access flow consists of 4 main steps:

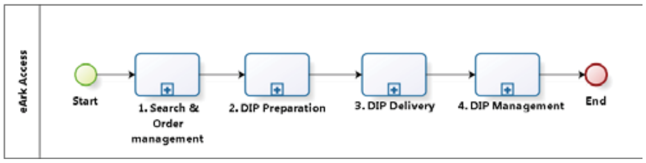


Figure 2. High-level access flow illustration

Each step is broken down in several sub-processes, an example of which is the DIP Delivery:

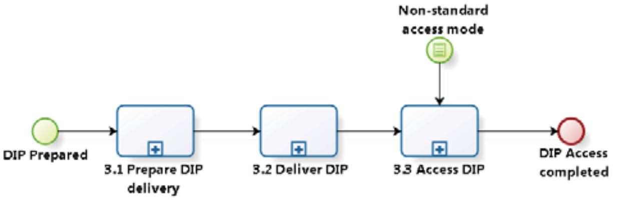


Figure 3. DIP Delivery

And each of these sub-processes is equally decomposed one more time, but those illustrations are too detailed to be inserted into this article.

In addition to the illustrations and textual descriptions of each process step, considerations were provided regarding the *product* context, which details in which environment do the products need to function and what the relationships are to native products and systems; the *assumptions* that represent the lowest common denominators that the system needs to heed, for example equipment availability, user expertise, and legal requirements.; and the *dependencies* that take into account relationships to other components and formats adopted or developed by the E-ARK project.

3.2.2 Identification of Use Cases

The identification and the description of the generic process steps enabled the creation of use cases, which have

1. served as communication platforms between archivists and developers and thus been used to facilitate the creation of an agile development environment where iterations rectify potential misconceptions;
2. informed the requirements of the access tools and the DIP format.

The use cases that were identified are as follows:

- Search in descriptive metadata and data
- Create initial order
- Validate order
- Check DIP availability and suitability
- Create DIP from AIP
- Modify DIP
    - Modify DIP for databases
    - Modify DIP for records
    - Modify DIP for GIS content
- Negotiate / prepare delivery method
- Provide access rights
- Notify end user
- Search in Database & ERMS
    - Search in database with Google functionality
    - Search with existing search forms
    - Search with SQL / DBMS functionality
    - Search with combination of google search and SQL / DBMS functionality
- Search in single records
- Search in GIS data
- Analyze with OLAP
- Deletion or maintenance of DIP
- Non procedural (generic) tasks
    - Update roles and users
    - Update access restrictions
    - Check logs
    - Log in

The use-cases are subject to change, and should be, throughout the whole agile development process.

3.2.3 *Acceptance Criteria, Constraints and Functional Requirements*

The use cases were enriched with acceptance criteria and constraints that in essence define quality goals (how will the product satisfy the user?) and boundaries (how is the product limited by external circumstances?) of the access services.

From each use case it was also possible to derive the functional requirements that were then matched to the functional requirements identified by the bottom-up approach described previously.

The bottom-up and the top-down approach have thus been used as complementary information processing methods and were adopted to secure a comprehensive understanding of the field of research at hand; and that all requirements were taken into account.

4. THE EARK DIP

4.1 A Common Specification for Information Packages

A set of common principles for all Information Packages in E-ARK have been developed to ensure consistency and coherence across IP formats. This framework called the Common Specification [4] outlines the structure of IPs, defines common metadata profiles, naming conventions and other matters that need identical handling across IP formats. The Common Specification makes up the core specification, but is amended and enriched for ingest, archival storage and access purposes in SIP, AIP and DIP formats respectively.

Information Packages are wrapped and described by a METS file. A specific E-ARK METS profile has been developed that defines core set of mandatory metadata and optional metadata. Widely used metadata standards, e.g. PREMIS and, are incorporated and used for their specific purposes (i.e. respectively preservation metadata and encodings for the finding aid).

4.2 The DIP Format

The purpose of the E-ARK DIP format is to create a format which is as standardized as possible and which observes technical, legal, user and other identified requirements. It is primarily an exchange format used for access purposes. The DIP is built on the principles from the Common Specification, but is extended to fit the specific purpose of access.

4.2.2 *The DIP data model*

The IP data model defined in the Common Specification is depicted below, and replicated in the DIP:

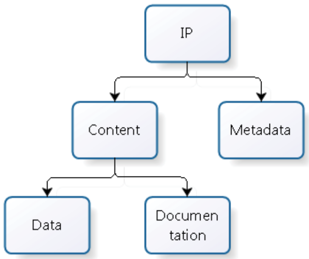


Figure 4. Common data model for Information Packages in E-ARK

As can be seen from the data model the DIP has a logical spilt between metadata and content, and content is further split into data and data-documentation.

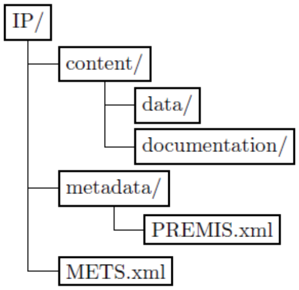


Figure 5. Common structure for Information Packages in E-ARK

The logical split is reflected in the Information Package structure (figure 5) where Information Packages are split into content and metadata at the root level of the package. The top level document of the IP is a METS document which describes the structure and encapsulates different types of metadata about the digital objects and inter-related metadata entities of the package. A common E-ARK METS profile is used across all E-ARK Information Packages. It has a basic set of mandatory elements that are common across all three types of IPs, but it can also be used to mark up metadata specific to SIPs, AIPs and DIPs respectively.

The E-ARK DIP format uses PREMIS to capture information targeted at supporting the digital preservation process. This means for example that the E-ARK PREMIS profile will capture preservation events pertaining to migrations. The conversion of a file into another format (named a ‘representation’ in E-ARK) will thus be documented in PREMIS. Another example is that E-ARK will make it optional to use PREMIS to capture relationships between an intellectual object or a representation to the documentation documents that are relevant for them. E-ARK will most probably adopt the new PREMIS 3.0[11].

I addition to the METS and PREMIS the DIP may include other metadata. This can be both structured metadata, including for example administrative-, preservation- and descriptive metadata in XML-format, and unstructured metadata, which could be scanned documents such as a user manual giving instructions about how the archived system was used when in production, classification schemas or filing plans, etc. It is up to each archive to decide what additional metadata and metadata files are included in the DIP. The E-ARK pilots will showcase different uses of the DIP format where it is adapted to local needs, specific content data types, and implementations.

The way “Content” is structured and documented in the DIP is not elaborated in this article because it has not been finalized at the time of writing. Detailed content type specifications will be developed for each of the content types in scope of EARK. The content type specifications will build on existing work e.g. SIARD2.0[12] will be used for databases and whole IT-systems.

[11] http://www.loc.gov/standards/premis/v3/index.html

[12] http://www.eark-project.com/news/29-siardfeedback

4.2.3 *Metadata in the DIP*

The majority of metadata in the DIP will be inherited from the SIP and the AIP and build on existing metadata files. In addition the DIP will contain DIP specific metadata, which for example could be information about which tool will be used to display the DIP in, i.e. "Rendering information" or supplementary authenticity information generated in the AIP-DIP process. Rights metadata are naturally also an important part of the DIP specific metadata, even though these can be embedded in the SIP and AIP profiles if needed. .

The METS file in the DIP will use the profile developed for the Common Specification, but as a DIP specific version of the profile holding DIP specific metadata elements. This means that if the "IP type" is set to "DIP" in the METS file, the elements in the file will accordingly be the ones relevant for access.

4.2.4 *DIP Specific Metadata*

The DIP specific metadata that are added in the DIP are divided into the following categories:

- Access rights
- User roles and permissions
- Rendering information, specifying for example that geo-data is to be rendered by QGIS[13]
- DIP status (there are three: One for when the DIP has been created (DIP$_0$); one when it has been prepared for the user (DIP$_u$); one for when it has been assigned to the DIP permanent storage (DIP$_p$))
- Dissemination notes and metadata enrichments made in the dissemination process
- Supplementary authenticity metadata (needed if for example multiple AIPs make up one DIP or if adjustments have been made to the DIP)

For each of the above categories specific metadata elements have been identified from existing standards and examinations of user needs. The metadata elements are described in-depth together with information about their datatypes, occurrence and whether or not they are mandatory.

4.2.5 *Access Related Metadata That Will Not Be in the DIP*

Not all Access related metadata should be included in the DIP. The dissemination process will depend on and generate other metadata than those inside the DIP. This can for example be metadata that the archives use to administer the DIPs and the dissemination process, or information about who has accessed a DIP and when. The purposes of them can be multiple, one of which is statistical.

These metadata types do not belong in the DIP, and it is up to each local archive to decide whether to keep them and where they should keep them (e.g. in their data management system).

4.2.6 *Data Formats in the DIP*

The formats of the data content types are not going to change whether they reside in the SIP, AIP or DIP.

As already mentioned they have not yet been entirely specified. The data content types that the E-ARK project will focus on are:

1. Single records, e.g. from ERMS (e.g. PDF, TIFF)
2. Databases (in SIARD 2.0 format)
3. Geo-data (in GML format)
4. Datasets for data mining (in OLAP cubes)

When the DIP has been created and exported to a staging area, it needs to be rendered to a viewer. If it's a database, it could for example to be loaded into a Database Management System (DBMS) and displayed in a Graphical User Interface (GUI), which is put on top of this DBMS.

5. ACCESS TOOLS

The access tools that will be developed represent all the components necessary to establish a fully functioning digital OAIS archive. This is the so-called "reference implementation" and will consist of open source code, which is downloadable from the source code management platform, GitHub[14]. The whole reference implementation can be downloaded and implemented, or just the desired components.

If an archive decides to download and install a single component, the open source code as well as a series of textual guidelines will facilitate the installation process. However, integration code will be necessary to develop as things like storage adapters are not within the scope of the E-ARK project.

Even though the intention is to offer a comprehensive digital archival solution, it is not all components that will be developed by the E-ARK project; actually, most won't, since the project essentially will build on existing resources: Where solid open source components exist, these will be integrated into the reference implementation; and where there's no urgent need for a component in the archival community, because every archive already has it, only basic functionality will be developed – this goes for example for the Finding Aid or the archival catalogue component.

As depicted in *Figure 2 - High-level access flow illustration* the E-ARK access system consists of four high-level processes for which four main software components will be developed. These will be made up of a yet undefined number of software modules.

The first one is the Search and Order Management component. This component consists of two main modules. The first one, Search and Select Information Objects, allows a consumer (cf. OAIS) to make searches in both data (the AIPs are loaded into a distributed storage (HDFS[15] and indexed in the Lily[16] repository) and metadata, using an open source Finding Aid yet to be decided upon. If an information object is not directly accessible (e.g. for reasons of access restrictions), the Manage Order module allows the archivist to validate or reject the order.

In case green light is given, the DIP Preparation component provides the appropriate DIP, either by fetching it in the DIP storage, or by creating it from parts of an AIP, one AIP or several AIPs. Since the IP formats are fashioned by the E-ARK project, it is necessary also to develop an AIP-DIP module that can access and transform the AIP(s) into a DIP. Another module in the DIP

Preparation component allows for modifying the DIP, if this is needed for reasons of for example anonymization of sensitive data.

[13] QGIS http://www.qgis.org/en/site/

[14] GitHub https://github.com/eark-project

[15] HDFS http://hadoop.apache.org/

[16] Lily http://www.lilyproject.org/lily/index.html

Once the DIP is prepared, it is sent into the DIP Delivery component. The DIP Delivery component notifies the consumer and choses the appropriate delivery scenario. Delivery scenarios depend on the nature of the content of the DIPs – the so-called data content types. The delivery scenario of for example a database can be executed in two ways: The priority will be to load the content of the database (which is wrapped in a SIARD 2.0 file) into a Database Management System (DBMS), which has an E-ARK built GUI sitting on top of it. This access scenario will require that users possess knowledge about SQL-queries in order to perform searches, but these searches will be powerful. If E-ARK resources allow for it, the second scenario for delivering a DIP database will be a NoSQL solution: The delivery module retrieves the database structure and data and sends them to an index engine (e.g. the de facto standard Lucene[17]).This allows for a far more user-friendly 'Google' like search, where the technical user requirements are minimal. The downside of this NoSQL scenario is however that the "Google" search method does not at all achieve the same level of pertinence of the results as queries made in a running DBMS do.

The last module – the DIP Management module – closes the process by sending the DIP to the DIP storage or deleting it.

The components, modules and storage areas used in the access system described above could look something like what is outlined below but not everything is settled at the time of writing.

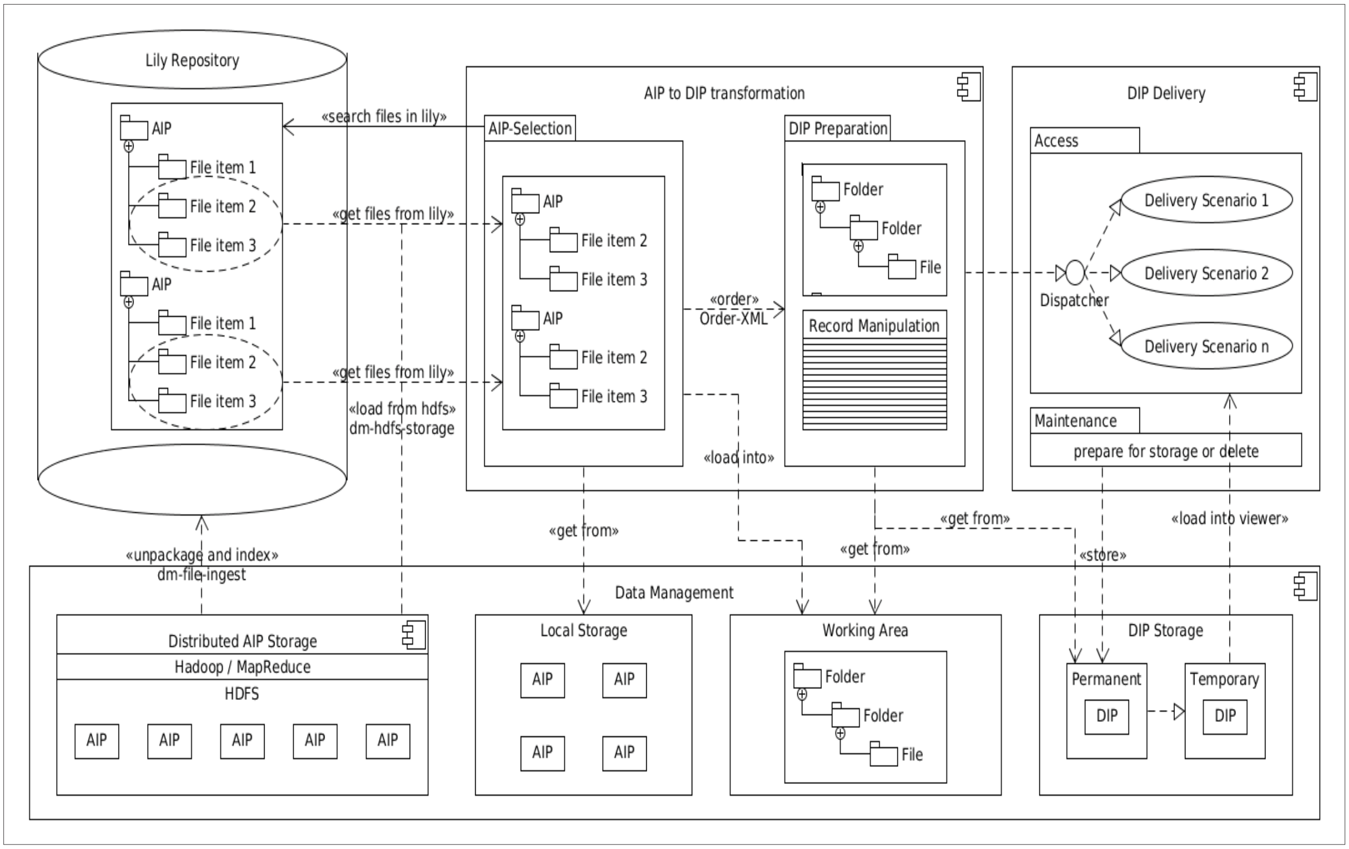


Figure 6. Overview over access components and modules

[17] Lucene https://lucene.apache.org/

6. PROOF OF CONCEPT: THE PILOTS

Ensuring that outcomes of the E-ARK have practical use and meet the needs of stakeholders and potential takers is crucial. To this end extensive piloting will be carried out.

A reference implementation of all components comprising an end-to-end solution will be hosted at the Austrian Institute of Technology (AIT). This will showcase the E-ARK end-to-end methodology for digital archiving.

In addition to the reference implementation seven pilot sites will test and implement E-ARK components. As can be seen from table 1, five pilots will focus on testing the access components and one will incorporate parts of the access components.

A real life local implementation of an end-to-end solution will be tested at Estonian Business Archives (Pilot 4).As an institution not directly involved in the E-ARK project the Estonian Business Archives will test products from a point of view of an institution wishing to implement a complete solution. Since this is a local implementation it will not necessarily include all E-ARK components but just the ones found suitable to meet the need in that particular situation.

Equally important to the full end-to-end pilot is the piloting of single E-ARK components. These will test the ability of E-ARK components to be implemented into existing environments. Testing the ‘plug-and-play’ aspect is vital because a major part of the archives in scope of E-ARK outcomes will already have a digital archiving solution in place and only wish to implement a subset of specific components.

Four pilots will test access components in combination with existing digital archiving environments. The focus of each pilot is different and different access components will be tested.

National Archives of Estonia will provide seamless access to public records (Pilot 3). National Archives of Slovenia will provide access to spatial data (Pilot 5). KEEP SOLUTIONS will load database into a DBMS for access purposes (Pilot 6). National Archives of Hungary to provide access to databases (Pilot 7). The Danish National Archives (Pilot 1) will test database access components in parallel with the pilot.

The extensive piloting and testing in multiple countries, technical environments and archival practices will ensure that E-ARK outcomes will in fact be scalable, robust and capable of meeting the needs of diverse organisations.

Table 1. Data Management and Access pilots of the E-ARK project

Full-scale Pilot		**Data Management**	**Access**
Pilot 1	SIP creation of relational databases (Danish National Archives)		🟧
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 geo-data (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)		🟦

🟦	Focus of the pilot
🟧	Elements also used/tried within the pilot

The pilots will start in late 2015 and continue throughout 2016 and thus run for about a third of the project time.

7. NEXT STEPS AND SUSTAINABILITY

At the time of writing the E-ARK DIP *draft* format was about to be handed over to the European Commission as an official deliverable of the project.

The next steps consist in establishing an environment where developers and archivists iteratively can enhance the DIP format and the requirements specification for the tools to be developed.

The last two tasks of the Work Package on Access will focus on respectively the development of an AIP-DIP transformation component; and the development of Search, Access, and Display Interfaces. The tools will handle single records, geo-data, databases, ERMS, and showcase data mining possibilities via OLAP cubes.

The E-ARK project believes that there is an interest for take-up of the tools that will be developed.

First of all, there’s a flagrant need for handling databases and EDRM systems in a standardized way, let alone at all. Every country has increasingly digitized administrations, which all use both. However, only a few national archives ingest databases on a scale that can keep the pace up with the public authorities’ use of these systems, and even fewer give access to them. There’s thus a growing need for ERMS and database archiving in Europe. Secondly, there is also a growing understanding of the fact that the only way of giving value to archival records is to make them accessible: Dark Archives are in nobody’s interest, and especially researchers and the public authorities themselves seek more and more frequently access to these records. And if national and local archives can display increased use of the records they hold, for example facilitated via E-ARK methodology and tools, increased funding cannot be too far away.

Thirdly, the E-ARK quantitative and qualitative interviews showed that user friendly tools as well as tools that help exploiting the information that lies within the IP’s, are much sought for. If E-ARK can fashion tools that respond to modern users’ expectations, there’s a good chance that they will be endorsed internationally.

The fourth consideration that will help increase take-up of E-ARK methodology is the ambition that the common IP format which is developed and operationalized by the E-ARK tools actually becomes the de facto standard of international archiving. Not only will it help the exchange of information packages and standardize the search for them and within them, but it will also reduce the number of tools needed in the archival community, and thus their development and maintenance cost.

Backing up the fourth consideration is a fifth one, which is all about that the fact that the tools are open source, and available on GitHub from the reference implementation. They will be accompanied by guidelines, but of course technical IT knowhow is indispensable for installing them.

Lastly, the pilot sites will integrate and use the tools, proving their worth; the tools will actually be running out there, before the E-ARK project ends, and be ready for direct implementation in an archive near you by the first months of 2017.

8. CONCLUSION

The E-ARK consortium is in the process of developing requirements specifications for the SIP, AIP and DIP formats as well as for the tools that will process these formats. These are based on thorough examinations of user needs, of the existing landscape of digital archival solutions in Europe, and of a series of other requirements that are relevant for the development of the E-ARK methodology, e.g. legislative and tools' requirements. The DIP format that results from these investigations is also to a very high degree based on existing standards, such as METS and PREMIS. In order to do a reality check regarding both the format and the tools, the E-ARK project envisages validation at pilot sites, which will prove the concepts. Measures for sustainability and up-take of products of the E-ARK tools and formats include an open source reference implementation holding independent software modules that can be downloaded and plugged in.

9. ACKNOWLEDGMENTS

The authors wish to thank Sven Schlarb, Austrian Institute of Technology (AIT) and the rest of the E-ARK group for their contributions.

10. REFERENCES

[1] Billenness, C. S. G., Anderson, D. and Johansen, K. H. E. 2014: E-ARK project – best practice survey results on archiving of digital material *In the proceeding of iPRES 2014 (Melbourne, Australia (6-10 October), 337-339* https://www.nla.gov.au/sites/default/files/ipres2014-proceedings-version_1.pdf

[2] Johansen, K. H. E., Thirifays, A., Randmäe, P., Škoflanec, J., Delve, J. and Anderson, D. 2014: Gap Report Between Requirements for Access and Current Access Solutions http://eark-project.com/resources/project-deliverables/3-d51-e-ark-gap-report/file

[3] Reference model for an Open Archival Information System (OAIS), ISO 14721:2012: https://www.iso.org/obp/ui/#iso:std:iso:14721:ed-2:v1:en:sec:3

[4] Common Specifications for IP's in the project E-ARK [This is an internal E-ARK project document; however it will be made available upon request. Contact one of the authors of this article to make a request.]

[5] Bergin, M. B., 2013: Sabbatical Report: Summary of Survey Results on Digital Preservation Practices at 148 Institutions, University of Massachusetts, Amherest: http://works.bepress.com/cgi/viewcontent.cgi?article=1012&context=meghan_banach

[6] Kristmar, K. V. 2012: Common challenges, different strategies", EBNA, 29 May 2012, Copenhagen http://www.sa.dk/content/us/about_us/danish_national_archives/25th_european_board_of_national_archivists_conference/presentations

[7] Ruusalepp, R. & Dobreva, M. 2012: Digital Preservation Services: State of the Art Analysis, Digital Cultural Heritage Network, DC-NET: http://www.dc-net.org/index.php?en/201/publications

Preserving the Fruit of Our Labor: Establishing Digital Preservation Policies and Strategies at the University of Houston Libraries

Santi Thompson, Annie Wu, Drew Krewer, Mary Manning, Rob Spragg
University of Houston
4333 University Drive
Houston, TX 77204-2000
1-713-743-9678
sathompson3@uh.edu; awu@uh.edu; ajkrewer@uh.edu; mmmanning@uh.edu; rspragg@uh.edu

ABSTRACT

To develop a comprehensive digital preservation program for maintaining long-term access to the Libraries' digital assets and align our practices with national standards and guidelines, the University of Houston (UH) Libraries formed the Digital Preservation Task Force (DPTF) to assess previous digital preservation practices and make recommendations on future efforts. This paper outlines the methodology used, including the task force's use of existing models and evaluation criteria, to successfully generate new policies and select Archivematica as our system to process and preserve our digital assets. It concludes with recommended strategies for the implementation of the policies and preservation operations.

General Terms

Institutional opportunities and challenges; Preservation strategies and workflows

Keywords

Digital preservation policy; System evaluation; Archivematica

1. INTRODUCTION

Creating, acquiring, preserving, and making accessible digitized and born digital content has been a major initiative of the University of Houston (UH) Libraries since the founding of the UH Digital Library in 2009. By the summer of 2014, UH Libraries had accumulated ten terabytes of digitized and born-digital content from UH Special Collections and the UH Digital Library.

UH Libraries established many of its digital preservation strategies and techniques for digitized materials within a year of creating the UH Digital Library in 2009. In their 2011 paper, "Implementing METS, MIX, and DC for Sustaining Digital Preservation at the University of Houston Libraries," Mingyu Chen and Michele Reilly outlined the original approach to digital preservation. The authors described a process that relied on a series of tools, including CONTENTdm export functions, JHOVE, and 7train, to generate descriptive and technical metadata in a METS wrapper [1]. Additionally, the article mentioned how UH Libraries was experimenting with the Texas Digital Library (TDL) to create additional storage locations for digital objects through a cooperative model including the Texas Advanced Computing Center (TACC) [1].

Over time, limitations to this model emerged. While it was critical to capture technical metadata, focusing exclusively on MIX metadata prevented the capture of technical information for other popular file formats, including audio, video, and datasets. The assembled tools also had no way of actively recording and tracking preservation events, through PREMIS metadata or any other mechanism. Perhaps the most important limitation, the existing tools and infrastructure had no formal digital preservation policy guiding current or future practices.

In response to these limitations, as well as to inconsistent practices around digital preservation, UH Libraries formed the DPTF in May 2014. The libraries charged the group with establishing a digital preservation policy and identifying strategies, actions, and tools needed to sustain long-term access to digital objects maintained by the libraries. Along the way, the DPTF combined existing research and evaluation criteria on digital preservation in new ways to generate robust policies and identify a new system that will enact these policies.

2. METHODOLOGY

The DPTF constructed its activities around its charge, which called on the group to:

- Define the policy's scope and levels of preservation
- Articulate digital preservation priorities by outlining current practices, identifying preservation gaps and areas for improvement, and establishing goals to address gaps
- Determine the tools, infrastructure, and other resources needed to address unmet needs and to sustain preservation activities in the future
- Align priorities with digital preservation standards, best practices, and TDL storage services
- Recommended roles, responsibilities, and next steps for implementing the strategy and policy

The DPTF launched parallel actions to fulfill its charge: policy development and system evaluation.

2.1 Policy Development

One activity focused on creating digital preservation policies for content entering into repositories. Policy creation required the group to study formal preservation frameworks, models, and strategies; compare UH Libraries' current practices with best practices from other libraries and archives; and craft new digital preservation policies that accounted for current technology and future resources.

The Action Plan for Developing a Digital Preservation Program (a toolkit distributed to participants in Cornell University's and The Massachusetts Institute of Technology (MIT)'s Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems workshop) served as the primary tool used for policy creation. Conforming to the Open Archival Information System (OAIS) Reference Model and the Trusted Digital Repository guidelines, the *Action Plan* walks institutions through the process of establishing a high-level framework, creating policies and procedures, building technological infrastructure, and addressing resources needed to sustain a digital preservation program for the long term. The document also includes policies and procedures from other institutions (some of which have used the *Action Plan*) [2].

To create the formal policies that inform UH Libraries digital preservation practices, the group performed the following activities:

- Selected and studied *Action Plan for Developing a Digital Preservation Program* to construct digital preservation policies
- Drafted high-level policy framework
- Outlined roles and responsibilities for internal and external stakeholders
- Defined digital assets including digitization quality and metadata specifications; collection selection, acquisition policies, and procedures; and access and use policies
- Identified and described key functional entities for the digital preservation system, including ingest, archival storage, preservation planning and administration, and access
- Drafted potential start-up and ongoing costs for digital preservation at UH Libraries

2.2 System Evaluation

Complementing policy creation, the task force also focused on evaluating software that UH Libraries will operate to fulfill the requirements of the digital preservation policy. The group reviewed research conducted by the Preserving (Digital) Objects with Restricted Resources (POWRR) project, including their Tool Grid and white paper "From Theory to Action: 'Good Enough' Digital Preservation Solutions for Under-Resourced Cultural Heritage Institutions." These resources provided valuable information on the capabilities and functionalities of over 60 tools and systems used for digital preservation activities [3] [4]. The DPTF used POWRR data to narrow potential tools to three for testing: Archivematica, Preservica, and Rosetta. After participating in demos on all three tools, the group elected to test Archivematica based on existing human and financial resources and in-house technological expertise. They chose not to test the other options because the task force found their costs to be prohibitive. Additionally, the proprietary software and structural metadata associated with the other two platforms were not based on open standards. To evaluate Archivematica, the task force adapted criteria developed by the National Library of Medicine (NLM) and made available in their 2008 report "Recommendations on NLM Digital Repository Software" [5]. Evaluation creation focused on the system's ability to support key services and functions, including:

- File types, including legacy formats
- Versioning control
- Virus and fixity checks
- Specified metadata formats
- Audit trail functionality
- Error reporting
- Archival Information Package (AIP) creation
- Dissemination Information Package (DIP) creation
- AIP storage

Finally, as the group tested Archivematica, it generated a list of questions and presented them to Courtney Mumma, the U.S. and International Community Development officer for Artefactual, Inc. during an onsite consultation with UH Libraries.

3. RESULTS

3.1 UH Libraries' Digital Preservation Policy

Principles outlined in UH Libraries' Digital Preservation Policy include collaboration, partnerships, and technological innovation, all of which are rooted in UH Libraries core values as articulated in both the 2013-2016 Strategic Directions document and our institutional mission. The University of Houston supports scholarship, teaching, and learning. As more resources and services associated with these functions become digital, our responsibilities must expand to include the identification, stewardship, and preservation of designated digital content. Additionally, UH has legal, contractual, and consortial obligations to preserve digital content of local and national significance.

The UH Libraries Digital Preservation Policy consists of three main sections: Policy Framework, Policies and Procedures, and Technological Infrastructure.

3.1.1 Policy Framework

The Digital Preservation Policy Framework supports the missions of UH and is the highest level digital preservation policy document at UH Libraries. It makes explicit UH Libraries' commitment to preserving the digital assets in its collections through the development and evolution of a comprehensive digital preservation program. The framework reflects the goals defined in our institutional mission and contains references to other relevant UH Libraries policies and procedures. The audience for the framework includes librarians and staff of UH Libraries, digital content donors/depositors, funders, and users [2] [8]. Sections in the Digital Preservation Policy Framework address:

- Purpose
- Objectives
- Mandate
- Scope
- Challenges
- Principles
- Roles and Responsibilities
- Collaboration
- Selection and Acquisition
- Access and Use

While it is outside the scope of this short paper to address all of these sections in the policy framework, key sections are described. The Purpose section explains the function of the policy framework and how it relates to more granular policies and procedures developed for UH Libraries [6]. The objective section articulates that UH Libraries defines the primary goal of digital preservation activities as maintaining the ability to meaningfully access digital collection content overtime. The primary concern is preserving the ability to access the archival digital object from which derivative files may be created or re-created over time [6] [7] [9]. Mandates for digital preservation at UH Libraries are dictated by fulfilling organizational commitments, including complying with the charge of the DPTF, supporting scholarship through long-term preservation of resources, maintaining institutional memory through preserving institutional records, and meeting any outstanding legal, contractual, or consortial obligations [6] [9]. The Scope section broadly outlines which assets will be retained and managed by UH Libraries. These assets include:

- Digital versions of resources owned and reformatted by UH Libraries and that fall under the parameters of UH Libraries' Digital Collection Development Policy
- Unique born-digital resources that are part of UH Libraries' archival/manuscript collections and which are unlikely to be preserved anywhere else
- Any other content acquired or digitized by UH Libraries that falls under the parameters of UH Libraries' Digital Collection Development Policy

3.1.2 Policies and Procedures

This section describes digital preservation policies, procedures, roles, and responsibilities in greater detail than the policy framework. This section outlines requirements around digital assets, including recommended capture specifications for digital objects, preferred file formats supported by the digital preservation system, and stipulations around the acquisition, transfer, and access of content [2] [6]. Additionally, this section of the policy addresses personnel. It identifies internal and external stakeholders, the roles required by the program, and the specific individuals charged with filling the roles [2] [6] [9].

3.1.3 *Technological Infrastructure*

UH Libraries' Digital Preservation Policy outlines digital preservation system functions and requirements in greater detail than the policy framework [2] [6]. Specifically, it articulates:

- The rules and requirements for Submission Information Packages (SIPs), Archival Information Packages (AIPs), and Dissemination Information Packages (DIPs)
- The workflow for ingesting, updating, storing, and managing digital objects
- The metadata requirements upon ingest
- The strategic priorities for future digital preservation efforts, including risk management

Functional entities implemented in UH Libraries' digital preservation system, such as pre-ingest, ingest, archival storage, data management, administration, preservation planning, and access are OAIS compliant.

3.2 UH Libraries' Digital Preservation System

The DPTF recommends that UH Libraries adopt Archivematica as its digital preservation system. In addition to this local storage solution, the task force also recommends storing digital objects in the cloud through DuraCloud services provided by TDL.

Rooted in digital preservation best practices, Archivematica combines numerous digital preservation tools to facilitate the acquisition, processing, and storage of digital objects. As an open source digital preservation system, Archivematica is designed to be extensible; the growing and active developer community continues to expand the tools and functionality of the system. It is also being developed to interoperate with other important digital access and preservation platforms, including DuraCloud and ArchivesSpace.

Using the modified NLM evaluation criteria, the task force identified advantages and disadvantages of Archivematica as a system and its implementation at UH Libraries.

3.2.1 *Advantages of Archivematica*

- Complies with OAIS reference model
- Uses open source solutions to perform digital preservation activities.
- Supports the ingest of a wide array of file formats
- Automates digital preservation policies, such as format choices when normalizing
- Offers active user development community
- Supports versioning through the adoption of the Archival Information Collection (AIC)
- Records digital preservation events and places this information into METS record as PREMIS metadata
- Offers an intuitive user interface that makes it easy for administrators to customize rules, settings, and workflows as well as to track workflow in a transparent way
- Supports complex archival workflows with multiple users having access, if desired
- Integrates with other digital asset management systems, including CONTENTdm, ATOM, and DuraSpace
- Provides a no-cost system solution with a pay structure for software support and/or customized features

3.2.2 Disadvantages of Archivematica

- Challenges IT staff due to its modular microservices architecture; it is built out of individual tools and is not a "set it and forget it" platform.
- Restricts the ingest of descriptive metadata to CSV file or manual input
- Stores objects in one specified location
- Lacks functionality to self-heal corrupted and/or damaged objects
- Limits the roles for users and administrators
- Lacks robust reporting and notification to assist with digital curation tasks

Despite the disadvantages (which could change over time because the system is actively developed), the task force believes that Archivematica offers a good balance of system functionality, future expansion, and ongoing sustainable costs. The task force will evaluate the disadvantages, prioritize them, and find partners to co-fund development solutions. Additionally, other groups, like DuraSpace, could address some of the identified deficiencies in the future.

To complement Archivematica, the DPTF recommends that UH Libraries' store copies of its content with DuraCloud, a cloud-based digital preservation solution. The task force selected DuraCloud because it can be synced directly with Archivematica, allowing for an automated delivery process. Additionally, DuraCloud is fully supported by TDL, which provides other critical services to UH Libraries.

4. IMPLEMENTATION STRATEGIES

With the completion of policy creation and system selection, implementing the newly established program will be the next step. The DPTF suggests that a new group, which we refer to as the Digital Preservation Team (DPT), be formed to create specific workflows that maximize Archivematica's ability to execute digital preservation policies. The creation of this team will allow engaged stakeholders to leverage their diverse knowledge and growing expertise in digital preservation in order to establish day-to-day workflows and procedures. DPT members should resolve several short-term priorities including:

- Training team members on the features and functionality of Archivematica
- Establishing workflows for digitized and born digital content that meet specifications for SIP, DIP, and AIC creation, dissemination, and storage as outlined in the UH Libraries Digital Preservation Policy
- Configuring system settings in Archivematica to automate aspects of the digital preservation policy

Additionally, the team should plan for long-term objectives, including:

- Collaborating with libraries stakeholders to identify and integrate areas where the digital preservation system and the new digital asset management system interoperate
- Advising the libraries on digital preservation policies related to materials that have no existing guidelines, such as electronic serials that are produced by the University and require preservation.
- Assessing and adapting workflows over time to increase efficiency and ensure compliance with policies

5. CONCLUSION

To date, the work of the DPTF has created a model that can inform the larger profession and has benefitted our local institution. The task force combined existing digital preservation research and evaluation criteria in new ways to generate robust policies and to identify a system to sustain these policies. DPTF members believe that sharing this information with external institutions could offer them an evaluation technique to draw upon when beginning the process of establishing digital preservation policies. Locally, the task force linked digital preservation issues with the mission of UH Libraries and aligned the libraries practices with national standards and guidelines, specifically OAIS and the requirements outlined by the Trusted Digital Repository model. By ensuring continued access to the libraries valuable and unique resources, we are protecting substantial institutional investments and supporting the University's goal to establish itself as a preeminent public research university in the 21st century.

6. REFERENCES

[1] Chen, M. and Reilly, M. 2011. Implementing METS, MIX, and DC for sustaining preservation at the University of Houston Libraries. *Journal of Library Metadata* 11 (2): 83-99. DOI= http://dx.doi.org/10.1080/19386389.2011.570662

[2] Kenney, A. and McGovern, N. (n.d.) *Action plan for developing a digital preservation program.* Cornell University and the Massachusetts Institute of Technology.

[3] Preserving (Digital) Objects with Restricted Resources (POWRR). Tool grid. Retrieved on 2015 April 16 from http://digitalpowrr.niu.edu/tool-grid/

[4] Schumacher, J., Thomas, L.M., VandeCreek, D., Erdman, S, Hancks, J., Haykal, A., Miner, M. Prud'homme, P., Spalenka, D. 2014. From theory to action: "Good Enough" digital preservation solutions for under-resourced cultural heritage institutions. A digital POWRR white paper. Retrieved on 2015 April 16 from http://commons.lib.niu.edu/bitstream/handle/10843/13610/FromTheoryToAction_POWRR_WhitePaper.pdf;jsessionid=314748CEF2C065FBE290B0B99A9E04AB?sequence=1

[5] National Library of Medicine (NLM) Digital Repository Evaluation Selection Working Group. 2008. Recommendations on NLM digital repository software. Retrieved on 2015 April 16 from http://www.nlm.nih.gov/digitalrepository/DRESWG-Report.pdf

[6] University of Houston Libraries Digital Preservation Task Force. 2015. University of Houston Libraries Digital Preservation Policy. University of Houston Libraries. Retrieved on 2015 November 15 from http://info.lib.uh.edu/sites/default/files/docs/DigPresPolForWeb_20150917.pdf

[7] National Library of Australia (NLA). 2013. Digital preservation policy, 4th edition. Retrieved on 2015 April 20 from https://www.nla.gov.au/policy-and-planning/digital-preservation-policy

[8] The Inter-University Consortium for Political and Social Research (ICPSR). 2012. ICPSR digital preservation policy framework. Retrieved on 2015 April 20 from http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/policies/dpp-framework.html

[9] The Ohio State University (OSU) Libraries. 2013 August. Digital preservation policy framework. Retrieved on 2015 April 20 from http://library.osu.edu/documents/SDIWG/Digital_Preservation_Policy_Framework.pdf

Preserving an Evolving Collection: “On-The-Fly” Solutions for *The Chora of Metaponto* Publication Series

Jessica Trelogan
Institute of Classical Archaeology
University of Texas at Austin
3925 W. Braker Lane
+1 (512) 232-9317
j.trelogan@austin.utexas.edu

Maria Esteva
Texas Advanced Computing Center
University of Texas at Austin
J.J. Pickle Research Campus
+1 (512) 475-9411
maria@tacc.utexas.edu

Lauren M. Jackson
Institute of Classical Archaeology
University of Texas at Austin
3925 W. Braker Lane
+1 (512) 232-9322
lmjackson@utexas.edu

ABSTRACT

As digital scholarship continues to transform research, so it changes the way we present and publish it. In archaeology, this has meant a transition from the traditional print monograph, representing the “definitive” interpretation of a site or landscape, to an online, open, and interactive model in which data collections have become central. Online representations of archaeological research must achieve transparency, exposing the connections between fieldwork and research methods, data objects, metadata, and derived conclusions. Accomplishing this often requires multiple platforms that can be burdensome to integrate and preserve. To address this, the Institute of Classical Archaeology and the Texas Advanced Computing Center have developed a “collection architecture” that integrates disparate and distributed cyberinfrastructure resources through a customized automated metadata platform, along with procedures for data presentation and preservation. The system supports “on-the-fly” data archiving and publication, as the collection is organized, shared, documented, analyzed, and distributed.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows.

Keywords

Archaeological data; database preservation; collection architecture.

1. INTRODUCTION

In archaeology, as in many disciplines, digital scholarship continues to transform the research process at every stage, from the collection of primary data on site, through post-excavation study and analysis, to the final interpretation and publication of results. A major effect of this transformation is the drive to publish full data collections in addition to print (or electronic) books. The printed monograph, traditionally considered the ultimate goal and the “definitive word” of any academic archaeological project, is giving way to an open, online, and interactive model that reflects a larger continuum of interpretation and reinterpretation. To represent and preserve archaeological research in this way, complex technical infrastructures and services are needed to support and provide fail-safes for data and multiple, simultaneous functions throughout a project’s lifecycle. Storage, access, analysis, presentation, and preservation must be managed in a non-static, non-linear fashion within which data evolve into a collection as research progresses. In this context, data curation happens *while research is ongoing*, rather than at the tail end of the project, as is often the case. Such data curation may be accomplished within a distributed computational environment, as researchers use storage, networking, database, and web publication services available across one or multiple institutions.

Ongoing data curation can be burdensome and costly, and, until recently, there has been little professional incentive to do it [1]. Facilitating long-term access to a project’s full set of primary data along with evidence for the processes of data collection, analysis, and interpretation promotes reproducibility and data reuse, but is not a trivial goal [2][3]. Whereas print publications end up in a library’s custody, in this new model, maintenance and preservation not only of a project’s data, but also of its mode of presentation falls, in many instances, to the research unit, requiring a post-custodial approach [4]. This is especially so when data publication requires more sophisticated technical resources than the average institutional repository can provide. This can include, as in the example we present here, web services and database and GIS technologies. Such requirements imply the backdrop of a solid infrastructure and a commitment to its long-term maintenance, and can require researchers to rethink data-intensive projects, reach out for expertise, cobble together adequate resources, and to implement more than one digital preservation strategy.

The Institute of Classical Archaeology (ICA) [5] is in the midst of a major program of study, synthesis, and publication related to long-standing field projects in the chora (countryside) of Metaponto [6]. For this initiative, a dispersed, multidisciplinary, and international team needs access to the legacy collection, a place to incorporate and share up-to-date versions of current work, a stable technical platform for managing data, and a space for continuing dialog throughout. With the Texas Advanced Computing Center (TACC) [7], which provides computational resources and expert data services to the University of Texas System and at the national level, we have implemented an infrastructure solution to accomplish those goals, while facilitating data curation tasks that will ensure the collection’s preservation. In addition to storage, preservation, and computational resources at TACC, we leverage file sharing services provided by the University’s Academic Technology

Support (ATS) group [8] and web services provided by Liberal Arts Instructional Technology Services (LAITS) [9], which hosts ICA's websites, including Wordpress-based digital companions to the print books (see below). We call this distributed infrastructure a "collection architecture." It integrates domain-specific technical resources and procedures customized to represent ICA's specific research processes and results.

ICA's collection is actively evolving simultaneously in different development stages of active research, publication, and archiving. Acknowledging that data in active projects are most vulnerable to disorganization and loss, and recognizing the importance of prompt archiving, access, and reuse, we consider preservation to be a constant activity that starts from the moment data are created and lasts throughout the collection's continuum.

2. THE ICA COLLECTION

ICA's data collection represents over forty years of research activities carried out since its establishment in 1974. Like any archaeological collection with such a long history, it reflects a dizzying number of technological, methodological, and theoretical changes that have influenced the field of archaeology and associated disciplines since the mid-1970s. It includes many types of data from a multitude of disciplines, from scans of analog photography, original drawings, and field notes to GIS data, born-digital imagery, full publications, and complex relational databases, each with its own set of methods, research questions, and technological requirements.

Currently ca. 5TB in size, the collection consists of data from more than twelve multi-year field projects in southern Italy and Ukraine, and a full range of associated specialist studies. It is growing rapidly as ICA's large physical archive is digitized and as new studies are conducted in support of the publication series. It is also riddled with duplication and redundancy [10], reflecting the recordkeeping habits and collected data silos from a huge, revolving team of people.

3. COLLECTION ARCHITECTURE

Over the course of the last six years, ICA and TACC developed the collection architecture presented here (Figure 1), which leverages existing storage, computing, cloud, and networking resources at the University of Texas at Austin [11][12]. The system enables data sharing and archiving "on-the-fly," as the collection is organized, documented, and analyzed during study and publication. These activities happen in parallel and behind the scenes in the collection architecture, which is distributed across major computational resources within the University. We have implemented services that include a GIS server and a set of web-based databases and Wordpress sites associated with each of ICA's archaeological projects. Metadata—extracted automatically where possible—fulfills data integration and preservation roles, and multiple preservation strategies assure data integrity and security throughout research stages and infrastructure components.

Mapped onto the collection architecture, an overview of our workflow is as follows. Messy legacy and new incoming data are first sorted by ICA research staff into broad categories in hierarchically labeled folders (the recordkeeping system), within a networked file share that functions as a staging area. These general categories (see Figure 2) provide basic descriptions, provenance, and context to data objects and help sift the collection into manageable chunks that relate to specific sites or specialist studies. Roughly organized data are then moved to a secure, geographically replicated storage resource (Corral with iRODS), where they are given unique identifiers. Thus, notably, data are archived at the outset, before further value is added to them through specialist study. From the archive, data objects are shared with the rest of the research team via web services through a web-based, domain-specific, GIS-enabled database (see ARK section, below). From here, the team studies the fully contextualized collection and adds further descriptions and connections as interpretations develop. The architecture allows the archaeological team to focus on research and publication activities, while metadata integration and preservation happens simultaneously in the background. To facilitate data sharing and to complement the print publication series, the Wordpress sites provide a guided entry point for unfamiliar users to navigate the data collection within the database. In addition, they provide access to original field notebooks and intermediary grey literature that cannot be presented in print and are beyond the scope of the database. Thus, each component of the architecture has a unique function, described in detail below, and all the data are preserved.

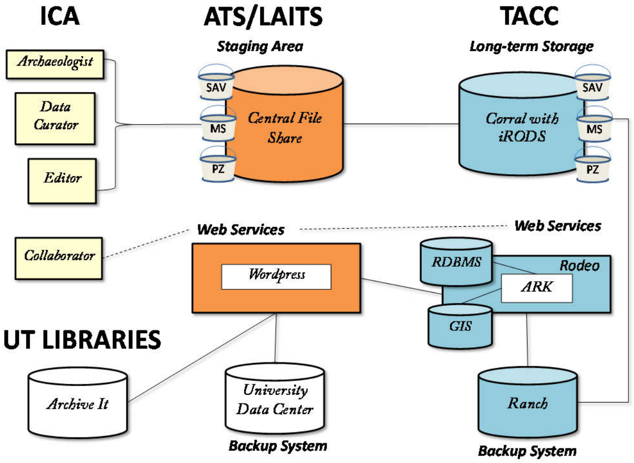


Figure 1. Collection architecture.

3.1 Staging Area and Recordkeeping System

Incoming data are moved into the collection architecture after being roughly sorted in the recordkeeping system within the central file share hosted by ATS. This recordkeeping system consists of a hierarchical file structure and naming conventions for various data types (Figure 2), which entail a neutral set of categories that are general enough to preserve vestiges of old recording methods and technologies, but also descriptive enough to make the collection navigable and reusable. The system is considered as a set of "big buckets" [13], the labels of which are used as descriptive metadata. In turn, the label terms have been mapped to the Dublin Core metadata standard [14] and are automatically extracted for every file as it moves from the file share to the storage resource, Corral with iRODS [15]. To preserve the integrity of the collection in terms of the fundamental archaeological principles of context and provenance, relationships between data objects and the sites and artifacts they represent are automatically captured from the recordkeeping system and recorded as metadata within Corral/iRODS.

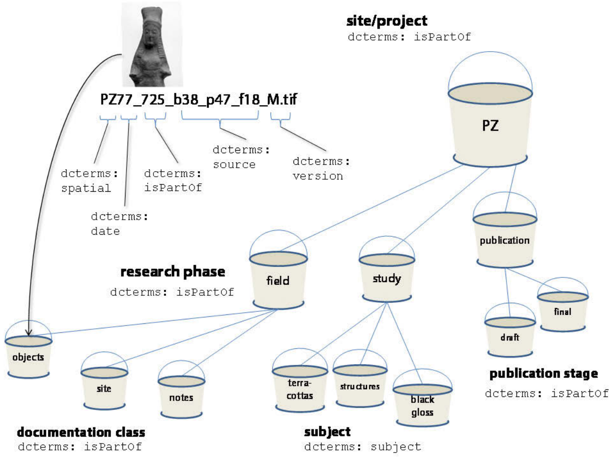


Figure 2. Recordkeeping system, implemented within the staging area and mirrored for long-term storage in Corral.

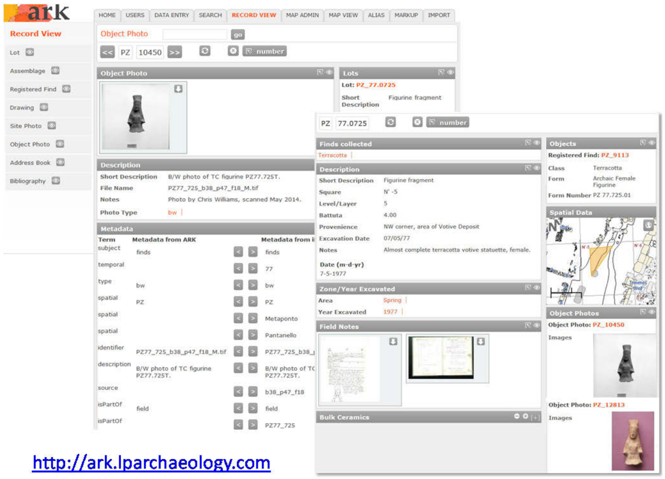
3.2 Corral with iRODS

Corral is a high performance storage system, geographically replicated, continuously monitored for security and failure, and available 24/7. It is part of the University of Texas System Research Cyberinfrastructure Initiative [16], which provides for its maintenance and expansion and subsidizes its cost. It is available to researchers in the UT System, who may have an initial allocation of 5TB of data for free. Corral uses iRODS as a data broker and rule engine, through which we enable—at ingest—automatic extraction of technical metadata along with descriptive metadata embedded in the file and recordkeeping system folder names. A checksum is also calculated for each file as part of the ingest process. This metadata gets registered in the iRODS iCAT metadata catalog for each file and is also formatted as a METS/Dublin Core/PREMIS file, stored along with the data object in Corral/iRODS. This automation provides documentation for every data object, its provenance, and relationships with other data objects and concepts *without any manual data entry* by the curators [15][12]. The data storage provides a long-term preservation solution for the primary data, which we refer to as the "archival instance" of the collection. Data are deposited here, documented, and preserved "on-the-fly," independent of their selection for further study or publication. The metadata gathered at this instance are preserved and integrated into ARK, the online database described below, to help users navigate the collection during study and make data reuse possible in the future. It also ensures the collection's integrity and helps reduce duplicated effort by providing a system of version control and tracking for each individual data object. This archival instance ensures the preservation of individual data objects and their metadata, acting as a fail-safe should any of the other components of the architecture (e.g., the online publication component) fail.

3.3 ARK (the Archaeological Recording Kit) and Rodeo

From the "archival instance" on Corral/iRODS, data objects and their Dublin-Core-mapped metadata are ingested into a web-based database built on the Archaeological Recording Kit (ARK), a pre-fabricated, open-source system [17] that required little extra investment in web development. ARK resides in Rodeo [18], TACC's cloud computing resource. Rodeo hosts a variety of databases and web services for the UT community in Virtual Machines (VM), allowing for fully customized computational environments and easy access to stored data from any location.

ARK's customizable structure and interface can be easily deployed for all of the varied archaeological projects[1] that are part of the Metaponto series—including excavation, survey, conservation projects, and museum exhibits—facilitating collaborative study and providing a central location for the international team to add details and make additional connections between related objects (Figure 3).



Object Photo

PZ 10450

Object Photo

Description

Short Description	B/W photo of TC figurine PZ77.725T.
File Name	PZ77_725_b38_p47_f18_M.tif
Notes	Photo by Chris Williams, scanned May 2014.
Photo Type	bw

Metadata

Term	Metadata from ARK	Metadata from i
subject	finds	finds
temporal		77
type	bw	bw
spatial	PZ	PZ
spatial		Metaponto
spatial		Pantanello
identifier	PZ77_725_b38_p47_f18_M.tif	PZ77_725_b38_
description	B/W photo of TC figurine PZ77.725T.	B/W photo of TC PZ77.725T.
source		b38_p47_f18
isPartOf	field	field
isPartOf		PZ77_725

Lots

Lot: PZ_77.0725

Short Description Figurine fragment

PZ 77.0725

Finds collected

Terracotta

Description

Short Description	Figurine fragment
Square	N' -5
Level/Layer	5
Battuta	4.00
Provenience	NW corner, area of Votive Deposit
Excavation Date	07/05/77
Notes	Almost complete terracotta votive statuette, female.

Date (m-d-yr)
7-5-1977

Zone/Year Excavated

Area	Spring
Year Excavated	1977

Field Notes

Bulk Ceramics

Objects

Registered Find: PZ_9113

Class	Terracotta
Form	Archaic Female Figurine
Form Number	PZ 77.725.01

Spatial Data

Object Photos

Object Photo: PZ_10450

Images

Object Photo: PZ_12813

Images

http://ark.lparchaeology.com

Figure 3. ARK screenshots: photograph stored in Corral/iRODS, metadata extracted from the recordkeeping system, the artifact's context within an excavation unit.

This part of the collection, which we refer to as the "study and presentation instance," also feeds directly into publication workflows by allowing the publication team direct access to artifact and site data as well as high-quality, original photographs and illustrations. More detailed metadata (dating, quantifications, typologies, etc.) can be entered here throughout study and pushed back to the persistent metadata storage system on Corral, so that at any point within the system, there is a full and up-to-date metadata record for each digital object. The evolving archive is thus constantly advancing, providing the basis for related studies, but is always secure. Once a project is complete and published, the ARK database is opened for public access and, via a persistent identifier (DOIs), the organized and fully-documented collection is ensured a permanent home for future access and further inquiry. For the presentation instances of *The Chora of Metaponto* series, we have configured one implementation of ARK per archaeological project. Each of these may have its own particular mode of presentation and contains its own set of data tables in ARK's database.

[1] "Projects" in this case may refer to any of ICA's excavation or surface survey campaigns. Each of these projects may contain more than one excavated site and may refer to more than one print monograph. ARK's flexibility means allows for a different configuration within each ARK instance, depending on the main unit of inquiry (e.g., the "site" in a surface survey, or the "stratigraphic unit" and "artifact" in an excavation).

3.4 Ranch

Ranch is TACC's massive tape-based, long-term storage system. Within our collection architecture, it is used as a high-reliability backup system for the study and publication instance of the collection. Here, we store routine backups of the ARK code base and custom configurations (see Preservation Strategies section below). Across Corral and Ranch, the entire collection architecture is replicated for high data availability and fault tolerance.

4. PUBLICATIONS INFRASTRUCTURE

4.1 Print Publications

The collection architecture functions as the data resource used during the publication process. Thus, specialist studies and interpretations, informed by and incorporated into ARK, either culminate in monographs within *The Chora of Metaponto* publication series, or appear as stand-alone articles, presentations, or grey literature reports. Since overall site interpretation relies upon the primary field documentation as well as dating and contextual information provided by multiple authors, constant access to full and up-to-date data via ARK expedites the creation of an accurate manuscript that reflects a cohesive understanding of the site or project.

4.2 Online Publications

A set of Wordpress-based websites serve as digital companions to the print publication series and as a portal to the data collections housed in ARK [19]. This service is hosted by LAITS as part of their remit to support faculty and staff research projects. The websites can either stand alone as a guided entry point to the data collection or to expand and complement interpretations presented in print. They also provide space to share full-resolution scans and transcripts of field notebooks, grey literature, and specialist reports related to the project. The blog platform's comment section permits immediate discussion and questions that can be directly connected to the original narrative in print, allowing the static interpretation to evolve with further research and input.

5. PRESERVATION STRATEGIES

Preservation is a key function requiring the implementation of more than one preservation strategy across the different infrastructure resources.

5.1 Integration of Data Objects and Metadata

All primary data objects are preserved in Corral/iRODS along with complete technical and descriptive metadata extracted at ingest. These are referred to via URIs within the ARK system, so that if users request a download of the original object, it comes directly from the archival instance on Corral along with its associated METS/PREMIS/DC record. When selected objects are called from the archive into ARK, a thumbnail is generated and descriptive metadata from the iRODS iCAT database populates basic information fields for that record. In turn, if extra descriptive metadata is added through the ARK interface during study, it is pushed back into the iCAT database. Thus, all the primary data and complete metadata are geographically replicated in case of failure of either component in the architecture.

5.2 Databases and Virtualization

While the complete Rodeo system that hosts the databases and the web code is backed up on a daily basis, such backups do not account for the specific workflows, data entry, and usage of individual projects. Thus, we implemented a customized database security and preservation strategy that could handle our ongoing publication production workflows and interfaces. To lower security risks, ARK's database is on one virtual machine, and its web code on another. By separating the database from the public access system we intended to avoid malicious breaches to the site's security. We created an automatic script to initiate daily SQL dumps of the ARK database tables, which are kept in a cascade: one a day for a week, one a week for a month, one a month for a year, and then one a year after that [20]. Additionally, virtualization was implemented as a preservation strategy in which the entire ARK database system running on the VM in Rodeo has a snapshot taken every night at 10 pm. This includes the accumulated SQL files that are produced earlier in the day. The resultant zip file is sent to the backup system in place on Ranch (see Ranch section, above) where we keep three days in a row and two months of backup files. This redundant approach avoids risks such as, for example, the unlikely corruption of files that could result from database writes happening at the same moment the database is snapshotted.

5.3 Wordpress Sites

LAITS provides cascading backups for files stored in the central file share and of the content of the Wordpress sites, with the latest versions discarded after 90 days. This type of backup is designed for disaster recovery as opposed to preservation of evolving interpretation. For this, we use the Archive IT service [22], sponsored by the UT Libraries, to archive snapshots of the Wordpress sites over time. At this time and until the publication is finalized we have scheduled monthly snapshot of the sites (e.g., http://wayback.archive-it.org/5446/20150508134828/
http://metaponto.la.utexas.edu/#).

6. CONCLUSIONS

Archaeological data are inherently vulnerable. Not only is excavation a destructive process, leaving the documentation the only remaining evidence of a site as it is uncovered, but archaeological collections can present serious data preservation challenges during and after a project. These collections tend to be accumulated and studied over decades, are especially large and complex, and reflect a huge range of technical sophistication.

In this project, we perceive preservation as an ongoing activity, which happens throughout the research process and continues well beyond a project's lifecycle into long-term maintenance of the published datasets. Data that are well organized, well documented, and authenticated from the beginning of the project are less vulnerable. We use a distributed set of diverse resources within which we are able to organize, describe, integrate and share data while archiving behind the scenes. In this system, raw, in-progress, and finalized data and publications are constantly secured using a variety of preservation strategies relevant to the different functions and technologies supporting the collection.

The solutions presented here have gone a long way toward streamlining ICA's publication and data sharing efforts and have ensured that a vulnerable collection is archived from the earliest stage possible. By leveraging existing University resources and expertise, the ICA team has been able focus on what it does best—archaeological research—and on enhancing the presentation of our results to provide more sophisticated interactive experiences to our target audiences. The next phase of our work will focus on issues of data reuse. The University of Texas Library supports the use of DOIs and ARKs (archival resource keys) [21], which we have begun minting for our data collections.

The administration and maintenance of the systems through TACC, ATS, and LAITS, are handled by people with the appropriate expertise. Nevertheless, implementing and maintaining this distributed infrastructure required extensive involvement and a learning curve for the domain expert data curators. For similar projects with large legacy collections in a push to publish a backlog of material, an “on-the-fly” approach like the one we present here can help alleviate the burden involved in making data comprehensible and reusable, while simultaneously preserving it as research progresses.

A major challenge that arises with the post-custodial approach adopted here, especially for grant-funded units like ICA, is to find an institution that can commit to maintain the fully functioning and dynamic set of ARK databases and the associated Wordpress sites for the long term. At the same time, thanks to this post-custodial approach, we know we have created a sustainable, well-documented platform that will make it easy to transfer once we do find such a host. Meanwhile, the metadata-ready archive can be deposited in an archaeological repository or at the UT Libraries as a static collection.

7. ACKNOWLEDGEMENTS

We would like to thank the Packard Humanities Institute (PHI) for enabling this research and for its continued generous support of *The Chora of Metaponto*, *The Chora of Croton,* and *Chersonesan Studies* publication series.

8. REFERENCES

[1] Kansa, E. C., and Kansa, S. W. 2011. Toward a Do-It-Yourself Cyberinfrastructure: Open Data, Incentives, and Reducing Costs and Complexities of Data Sharing. In *Archaeology 2.0: new approaches to communication and collaboration*. E. C. Kansa, S. W. Kansa, and E. Wattrall, Eds. 57–92. Los Angeles: Cotsen Institute of Archaeology Press.

[2] Borgman, C. L. 2012. The conundrum of sharing research data. *J. Am. Soc. Inf. Sci.* 63: 1059–78. DOI= http://dx.doi.org/10.1002/asi.22634.

[3] Frank, R. D., Yakel, E., and Faniel, I. M. 2015. Destruction/reconstruction: preservation of archaeological and zoological research data. *Journal of Archival Science* 2015-01-11: 1–27. DOI= http://dx.doi.org/10.1007/s10502-014-9238-9.

[4] Henry, L. J. 1998. Schellenberg in Cyberspace. *American Archivist* 61.2: 309–27.

[5] *Institute of Classical Archaeology (ICA)*. Accessed 10 August 2015. http://www.utexas.edu/cola/ica/.

[6] Carter, J. C, ed. 1998–2014. *The Chora of Metaponto*. Vols. 1–5. Austin: University of Texas Press.

[7] *Texas Advanced Computing Center (TACC)*. Accessed 13 August 2015. https://www.tacc.utexas.edu/home.

[8] *University of Texas at Austin Academic Technology Support (ATS)*. Accessed 15 September 2015. https://www.utexas.edu/transforming-ut/shared-services/ats.

[9] *College of Liberal Arts Instructional Technology Services (LAITS)*. The University of Texas at Austin. Accessed 15 September 2015. http://www.utexas.edu/cola/laits/.

[10] Arora, R., M. Esteva, and J. Trelogan. 2014. Leveraging High Performance Computing for Managing Large and Evolving Data Collections. *International Journal of Digital Curation* Vol. 9, No. 2: 17–27. DOI= http://dx.doi.org/10.2218/ijdc.v9i2.331.

[11] Esteva, M., Trelogan, J., Rabinowitz, A., Walling, D., and Pipkin, S. 2010. From the site to long-term preservation: a reflexive system to manage and archive digital archaeological data. In *Archiving 2010. Proceedings of the Archiving Conference, Vol. 7 (Den Haag, the Netherlands, June 1–4, 2010)*. 1–6.

[12] Kulasekaran, S., Trelogan, J., Esteva, M., and Johnson, M. 2014. Metadata Integration for an Archaeology Collection Architecture. In *Proceedings of the International Conference on Dublin Core and Metadata Applications (Austin, Texas, 8–11 October 2014)*. 53–63. Retrieved from http://dcpapers.dublincore.org/pubs/article/view/3702.

[13] Cisco, S. 2008. Big buckets for simplifying records retention schedules. *ARMA International’s Hot Topic 2008*: 3–6. Retrieved 9 May 2014 from http://www.emmettleahyaward.org/uploads/Big_Bucket_Theory.pdf.

[14] Dublin Core Metadata Initiative. DCMI Specifications. Accessed 20 April 2015. http://dublincore.org/specifications/.

[15] Walling, D., and Esteva, M. 2011. Automating the Extraction of Metadata from Archaeological Data using iRods Rules. *International Journal of Digital Curation* Vol. 6, No. 2: 253–64. DOI= http://dx.doi.org/10.2218/ijdc.v6i2.201.

[16] The University of Texas System Research Cyberinfrastructure (UTRC). Accessed 13 April 2015. http://www.utsystem.edu/offices/health-affairs/utrc/storage.

[17] Eve, S., and Hunt, G. 2008. ARK: a developmental framework for archaeological recording.” In *Layers of Perception: Proceedings of the 35th International Conference on Computer Applications and Quantitative Methods in Archaeology (CAA) (Berlin, Germany, April 2–6, 2007)*. A. Posluschny, K. Lambers, and I. Herzog, Eds. Kolloquien zur Vor- und Frühgeschichte Band 10. Bonn: Dr Rudolf Habelt GmbH. Retrieved from http://proceedings.caaconference.org/paper/09_eve_hunt_caa2007/.

[18] Rodeo. 2014. Retrieved 14 August 2014 from https://www.tacc.utexas.edu/resources/data-storage/#rodeo.

[19] Institute of Classical Archaeology. The Chora of Metaponto: a digital companion to the publication series. Accessed 20 April 2015. http://metaponto.la.utexas.edu.

[20] Preserving Relational Databases. 2015. Retrieved 20 April 2015. http://digital.humanities.ox.ac.uk/Support/PreservingDatabases.aspx.

[21] California Digital Library. EZID. Accessed 13 April 2015. http://ezid.cdlib.org.

[22] Archive-It -Web Archiving Services for Libraries and Archives. Retrieved 29 April 2015. https://archive-it.org/.

Functional Access to Forensic Disk Images in a Web Service

Kam Woods
UNC Chapel Hill
CB #3360, 100 Manning Hall
Chapel Hill, NC 27599-3360
919-962-8366
kamwoods@email.unc.edu

Christopher A. Lee
UNC Chapel Hill
CB #3360, 100 Manning Hall
Chapel Hill, NC 27599-3360
919-962-8366
callee@ils.unc.edu

Oleg Stobbe
University of Freiburg
Hermann-Herder Str. 10
79104 Freiburg, Germany
oleg.stobbe@rz.uni-freiburg.de

Thomas Liebetraut
University of Freiburg
Hermann-Herder Str. 10
79104 Freiburg, Germany
thomas.liebetraut@rz.uni-freiburg.de

Klaus Rechert
University of Freiburg
Hermann-Herder Str. 10
79104 Freiburg, Germany
klaus.rechert@rz.uni-freiburg.de

ABSTRACT

We describe a hybrid approach for access to digital objects contained within forensic disk images extracted from physical media. This approach includes the use of emulation-as-a-service (EaaS) to provide web-accessible virtual environments for materials that may not render or execute accurately on modern hardware and software, and the use of digital forensics software libraries to produce web-accessible file system views to support single-file access and provide visualizations of the file system.

General Terms

Frameworks for digital preservation; preservation strategies and workflows

Keywords

Emulation, access, digital forensics

1. INTRODUCTION

Support for meaningful use of digital objects often requires retention of the environment (or aspects of the environment) in which they were produced. This can help to reproduce significant properties of the digital objects [2], as well as reflecting essential contextual information.

For materials acquired on fixed and removable digital media, addressing this need begins with acquiring a complete disk image, which is a block-by-block copy of the disk's storage. No prior knowledge of the operation system (OS) or file system on the disk is required to perform the acquisition. Similarly, one can search for patterns within the bitstream (e.g. email addresses, credit card numbers, phone numbers) without necessarily knowing or having software support for the original OS or file system [1].

Analysis and description tasks can require mounting of the original file system. These include navigation of the files and folders; extraction of specific files or folders; extraction of file system metadata; and reporting the number and types of files on disk. Additional digital curation actions also require software that can recognize, access and render information from specific file formats. These include file characterization, validation, metadata extraction and visual inspection.

File systems and file formats are subject to obsolescence, and digital curation professionals often process born-digital materials that are not supported by contemporary computing environments. One response to this challenge is to install dedicated software (applications or complete operating systems) on the machine being used to process the materials, or to consolidate these tools into a specialized environment. An example of this is the BitCurator environment[1], a suite of open source digital forensics and data analysis tools to help collecting institutions (libraries, archives, and museums) process born-digital materials. This environment, developed through a series of grants from the Andrew W. Mellon Foundation, has been customized to work with many obsolete file systems and file types. It also contains software for the creation of forensic disk images; analysis of files and file systems; extraction of file system metadata; identification of sensitive information; and identification and removal of duplicate files.

There is always the possibility of acquiring disks with file systems and files that are not supported by the available tools. One also cannot assume that end users will be running specialized tools on their local machines. An alternative access strategy is emulation: enabling the user to boot and

[1] BitCurator, `http://www.bitcurator.net/`

interact with an original operating system, or attaching the disk as a secondary drive to an emulated environment typical of the era in which it was produced. Emulation-as-a-Service (EaaS) simplifies this process for end users by providing access to pre-configured emulation environments within a web browser.

We present an approach to accessing operating systems and file systems contained in disk images using both EaaS[2] and a dedicated web application to generate views into non-live file systems. For public (or semi-moderated) access, redaction of sensitive content is often required. We describe a traceable redaction workflow and implementation for restricted functional access to disk images, supporting different access levels depending on the requester's role.

2. RELATED WORK

Capture and analysis of disk images from fixed and removable media is a mainstay of digital forensics practice. The need to quickly analyze large quantities of digital information has led to the development of several modular open-source tools and platforms to parse file system contents and identify and analyze features of interest within the file systems [1].

The development of open-source digital forensics tools to manipulate common disk image file formats (along with tools to create and export from them) increases the attractiveness of digital forensics tools to collecting institutions. These include *libewf*, an open source library to create and manipulate files in the widely-used Expert Witness Format [4].

3. ACCESS WORKFLOW DESCRIPTION

For purposes of this discussion, we assume that one has already created a disk image along with a description of the disk's technical environment (e.g. source descriptions, size in bytes, file system(s) present). This information is used to make decisions about the access environment and enable preparation of any surrogate – for example, if there is data within the disk image that requires redaction.

3.1 Preparation

It is important to distinguish between two distinct access modes: interacting with the disk image as a bootable system disk (e.g. if the disk contains an operating system); and attaching the disk as a secondary disk to an emulated environment. In the latter case, no further measures are required. The former case requires a hardware generalization process that we describe below.

First, a description of the original hardware environment associated with the disk image is examined to identify the correct emulator configuration (or locate a similar emulated system prepared previously). Emulators typically provide only a limited selection of hardware system components – usually popular devices with broad driver support. For systems from the 1980s and 1990s, for example, common ISA-bus hardware devices with virtualization support include the Soundblaster 16 and AdLib sound cards, the NE2000 network adapter, and the Cirrus VGA graphics adapter.

To recreate a system associated with specific hardware, additional hardware drivers may need to be installed or replaced – at least if full functionality is required. This process may require certain changes to the disk image. These changes, however, must not be applied directly to the (forensic format) disk image, but have to be kept as a separate change-set, which supports tracking of (technical) modifications both on a block and file system level.

The result of the preparation process is a set of technical changes required to run on a generic emulated computer system. While the acquired image may be altered, this generalization process also comes with benefits: the machine setup is fully documented and understood, and hardware dependencies are explicit and can serve as a preservation and planning guide for emulating other systems in the future.

3.2 Redaction & Dissemination

Bootable system disks are more likely to contain items that require redaction, including personally identifying and sensitive information within documents explicitly produced by the original user(s), and other data retained via the normal operation of the operating system and file system.

As an example, Windows-based systems retain information corresponding to various user activities, including passwords (which may not be well encrypted in earlier versions of the OS), lists of recently viewed documents, devices that were attached to the original system, and - potentially - sensitive data including online credentials and encryption keys. Depending on the version of the OS used when the system was active, this information may appear in the Registry, in the hibernation file used for fast resume on system wakeup, and in unallocated areas of the disk.

In past publications, we have shown how open-source digital forensics tools such as Simson Garfinkel's *bulk_extractor* may be incorporated into archival workflows, allowing users preparing collections for access to redact both at the block level on disk, and to restrict access to individual files [3].

Rather than storing raw disk images, many collecting institutions are using forensic disk image formats, such as EWF, which can compress the data, as well as embedding integrity checks and various forms of metadata. EWF files cannot be redacted in-place without compromising consistency checks internal to the file format. One workaround is to export the raw data from the EWF image, redact the relevant blocks (or mount and redact specific files or directories), and create a new EWF file using the redacted raw image. There are cases when this approach may not be desirable, because it complicates the provenance record of the stored data.

Alternatively, features identified by *bulk_extractor* may be linked to individual file items and recorded in an annotated Digital Forensics XML (DFXML) file. Working with *libewf*, an open-source software library, it is possible to create a synthetic listing of the contents of the file system within an EWF file that elides any file or directory item within the DFXML file that is marked as containing restricted material.

For non-emulated access on the Web – viewing the contents

[2] bwFLA EaaS, `http://bw-fla.uni-freiburg.de`

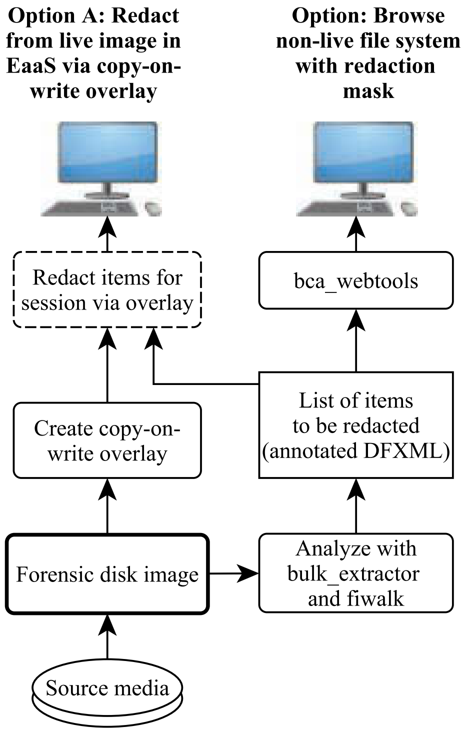

Option A: Redact from live image in EaaS via copy-on-write overlay
Option: Browse non-live file system with redaction mask
Redact items for session via overlay
bca_webtools
List of items to be redacted (annotated DFXML)
Create copy-on-write overlay
Forensic disk image
Analyze with bulk_extractor and fiwalk
Source media

Figure 1: Simplified workflow showing redaction options for emulation and browsing access.

of the file system in a simulated directory structure within a web page – this file may be consulted to determine whether a given file or directory should be presented to the user as a link. This effectively "blacklists" files containing restricted, sensitive, or private information from public access.

Creation of a "surrogate" EWF image using an exported, redacted raw image from the original EWF source is an obvious approach for facilitating restricted access. However, the storage and time requirements associated with creating altered copies of original disk images may be prohibitive – storage alone will effectively double unless the original image is discarded.

As an alternative, the blacklist annotations to the DFXML representation of the file system may be passed to the emulation tool to modify the file system immediately prior to user access, deleting file items and scrubbing unallocated areas prior to enabling user access. Some access options when providing redaction services for forensic disk images are shown in Figure 1.

3.3 Emulation-based Access

Web-based access to born-digital archival materials is often restricted to individual files that have been specifically selected for access. These files may be normalized (e.g. converting Microsoft Word to PDF/A), with the only context for the *original* environment being in the archival metadata that accompanies the file.

This can degrade the access experience in several ways. Executable content may not run on modern systems, or may depend on hardware that is not accommodated by (or simulated by) modern device drivers. Second, the user may be more interested in the original structure and organization of the content than the content itself. Finally, there may be features or limitations of the production environment (the bootable operating system) that are of interest with respect to their influence on the documents or media produced.

Emulated environments can provide a view of the original production environment, but have traditionally faced various hurdles, including lack of computing power on the end-user's system and lack of expertise in installing and configuring required software. Emulation-as-a-Service addresses these limitations by offloading the computational requirements to a hosted service and providing users with "one-click" access to bootable environments within a web browser. Emulation platforms such as QEMU provide access to a range of operating system environments and disk image formats, but support for formats most common in forensic disk imaging did not previously exist. In the following section, we describe a mechanism (including a novel QEMU block driver) to enable access to forensic disk images in EWF format.

3.4 Implementation

A disk image captured as an EWF file is effectively read-only. Any deliberate alterations to the content of the image will produce error warnings in libraries capable of reading the contents; these changes will cause embedded cyclic redundancy checks to fail.

To use the EWF image in an emulation setup, a writeable disk is required. As a first step we create a writeable *overlay file* that forwards read operations for any unmodified block to the original EWF file. Write operations are captured and only written to the overlay file. This process is known as *copy-on-write*. Subsequent reads of such modified blocks are severed from the overlay file. This mechanism allows data modifications to be stored separately, independent of the original digital object during an emulation session. This allows each digital object to be retained in its preserved, unmodified state. After an emulated session the overlay-file can either be discarded or kept for future use or analysis.

To achieve this we have implemented an EWF QEMU block driver to enable access using QEMU's disk image handling tools and to make use of QEMU's QCOW2 container format [3]. The QCOW2 format allows one to store all changed data blocks and the respective metadata for tracking these changes in a single file. To define where the original blocks (before copy-on-write) can be found, a *backing file* definition is used. QEMU's Block Driver API provides a continuous view on this QCOW2 container, transparently choosing either the backing file or the copy-on-write data structures as source.

As any block format is allowed in the backing file of a QCOW2 container, the backing file can itself be a QCOW2 container.

[3] The QCOW2 Image Format, `https://people.gnome.org/~markmc/qcow-image-format.html`, last access 4/8/15.

This allows "chaining" a series of modifications as copy-on-write files that only contain the actually modified data. One can use this feature to make individual changes to the original environment citable and accessible, for instance, to provide access to a disk's redacted version.

This overlay concept and its implementation does not depend on a specific emulator (such as QEMU). It may be adapted to work with any emulation platform that provides appropriate access. Listing 1 shows an example creating the overlay file `ewf-overlay.cow` using the backing file `ewf-demo.E01`.

Listing 1: Example creating a QCOW2 overlay on top of a EWF file.

```
qemu-img create \
 -f qcow2 \
 -o backing_file=ewf-demo.E01,backing_fmt=ewf \
 ewf-overlay.cow
```

To make use of the overlay file with an arbitrary emulator (including emulators with no native QCOW2 support) the raw payload needs to be exposed. This can be achieved by "fusing" the QCOW2 container to expose its raw content as a synthetic continuous file. Read and write operations are intercepted by the FUSE[4] file system layer and translated to appropriate QCOW2 read/write operations.

Listing 2 uses `qemu-fuse` to expose the raw disk image. The resulting file `raw-content/ewf-overlay.cow` contains the bit-exact copy of the original physical disk without any additional metadata added by the EWF format or QCOW2 container and therefore can be attached directly to an emulator.

Listing 2: Expose raw disk content using qemu-fuse

```
qemu-fuse ewf-overlay.cow raw-content/
```

Capturing changes at the lowest possible layer (the block layer) has specific technical advantages compared to higher layers (e.g. file system). First, this approach is independent of the hardware medium (disregarding vendor-specific storage areas on modern devices that have no effect on file system access), and does not depend on any operating system or file system encoded on the device. Second, the required metadata is simple and relatively easy to understand; reconstruction of the file is possible even without access to the original tools. Listing 3 shows metadata of an unmodified overlay file, with all blocks mapped to the (original) backing file.

Listing 3: The block mapping table before modification of the overlay file

```
Offset   Length      Mapped to   File
0        0x1f400000  0           ewf-demo.E01
```

Listing 4 shows metadata after modification[5]. Several blocks have been changed on the disk and are now mapped to the overlay file.

Listing 4: An excerpt from the block mapping table after modification of the overlay file

```
Offset    Length     Mapping   File
0         0x10000    0x60000   ewf-overlay.cow
0x10000   0x10000    0x10000   ewf-demo.E01
0x20000   0x10000    0x70000   ewf-overlay.cow
0x30000   0x10000    0x30000   ewf-demo.E01
0x40000   0x10000    0x50000   ewf-overlay.cow
0x50000   0x620000   0x50000   ewf-demo.E01
0x670000  0x10000    0x80000   ewf-overlay.cow
0x680000  0x1ed80000 0x680000  ewf-demo.E01
```

While critical to the implementation, these details are not visible to the end user. The user sees only an environment that can be navigated, modified, and otherwise interacted with, while the underlying disk image (the preservation object) remains unchanged.

4. USE CASES & EVALUATION

As outlined in the previous sections, we envision two basic use cases: a user browsing the file system of a forensic disk image via a web-interface, and a user interacting with a booted file system or secondary storage device via an emulated environment rendered within a web browser.

Both approaches support interaction with forensic disk images by providing access to the underlying file system(s) using existing open source libraries to read the contents of the disk image format.

4.1 Using an EWF Image as Boot Disk

To evaluate the capabilities of our tools and workflow we have chosen a real use case, demonstrating the image preparation process, i.e. a technical generalization to be used with an appropriate emulator.

The Vilem Flusser Archive owns a personal computer associated with the production of a software titled "Flusser-Hypertext". This computer contains a rare working copy of the software which is dependent on the obsolete authoring system HyperCard. The disk image has been acquired[6] from an Apple Mac Performa 630 containing a 270 MB IDE disk. The goal was to enable web-based access to the Flusser-Hypertext through the archive's web site.

Using the acquired disk image directly with an emulator failed. The original machine used a hardware-related extension (A/ROSE) that is not supported by the emulator used and prevented the system to start properly. A simple solution was to boot the system with all extensions disabled and to delete the A/ROSE extension file from the system's extensions folder. The result of this process is an overlay-file that is bootable and useable with an emulator. The overlay's file size is 823 KB and contains 7 changed blocks (block size was set to 1024 bytes). However, simply booting the (unmodified) file system results in 3 changed blocks.

4.2 Redaction for Public Access

The disk image is now fully functional in an emulation scenario. However, it is not yet suitable for public access. As sensitive private data was found on the disk image, these

[4] FUSE: Filesystem in Userspace, `http://fuse.sourceforge.net/`

[5] In this case a MS-DOS 6.2 system has been booted and a directory has been created on the disk

[6] The original acquisition was performed using `dd` without forensic tool support. We have reacquired the raw disk image as an EWF image.

files have to be removed, and a second overlay file has to be produced.

In case of the Flusser Mac the archive provided a list of files that are not suitable for public access. These files have been removed from the file system on a second overlay file, which is now accessible through the archive's web site[7] and citable. In general, the redacted version of the disk image is inextricably linked to the original image, such that any action of the redaction process can be audited.

4.3 Access using EaaS

Once an overlay file for public access has been created, it can be published using the EaaS framework. In a typical EaaS setup, the emulator runs either on a local computing cluster or using a cloud computing service (such as Google Cloud). Disk image storage, description, and publication is managed by the respective preservation institution.

To securely publish a redacted disk image, only a standard web server[8] is required. Ideally, the redacted overlay-file points to a local backing file that is not accessible through the web server. Properly implemented, this ensures that a user visiting the website cannot exploit a vulnerability of the server-side software to read data directly from the overlay (for example, to examine blocks from the original disk image that have been scrubbed, or files that have been "deleted" from the image via the overlay).

The EaaS service requires a `binding` configuration as part of the technical metadata, defining the data source' to be configured as an emulated machine's `drive`. Listing 5 shows an example of an EaaS configuration. If no redaction is required, or the emulator access is not public, a pointer to the EWF file (e.g. an HTTP link) in the bindings section is sufficient.

Listing 5: Metadata defining data resources and emulator medium mapping

```
[...]
 <drive>
  <data>binding://main_hdd</data>
  <iface>ide</iface>
  <bus>0</bus>
  <unit>0</unit>
  <type>disk</type>
  <boot>true</boot>
  <plugged>true</plugged>
 </drive>

 <binding id="main_hdd">
  <url>https://.../diskImage.pub</url>
  <access>cow</access>
 </binding>
[...]
```

In both cases these images can be cited (e.g. using HDL) and functionally accessed [9].

5. FUTURE WORK

Some of the uncertainty associated with handling disk images extracted from legacy media – particularly when they contain bootable operating systems – is derived from a lack of sufficient description of the technical environment in which they were produced. Providing guidelines for how those technical environments should be described is paramount in supporting the contextualization and generalization process. In future efforts, we intend to provide additional guidance on factors related to both the hardware and operating system that should (at a minimum) be recorded.

Some aspects of this process may be automated, particularly when working with operating systems such as Windows, OS X, and earlier version of the Macintosh operating system that record hardware characteristics in well-documented locations.

We also plan to further explore the relationships between EaaS and navigation of disk image file trees in a web browser [5]. We plan to examine options for creating richer, more unified interfaces to allow users to examine metadata related to disk images, search the contents of images prior to accessing them directly, and browse to EaaS instances from within existing archival access environments.

6. CONCLUSION

We have described a series of methods to provide web-based access to disk images captured in forensic formats; through an emulation system that can be accessed using a modern web browser, and by browsing views of the file system directly within a webpage. These approaches address an important need among collecting institutions: allowing users visiting their website to interact with legacy operating systems and file systems contained in disk images extracted from legacy media, without requiring them to install software or understand the technical details required to recreate a functional environment.

7. REFERENCES

[1] S. L. Garfinkel. Digital media triage with bulk data analysis and bulk extractor. *Computer Security*, 32(C):56–72, Feb. 2013.

[2] M. Hedstrom and C. A. Lee. Significant properties of digital objects: definitions, applications, implications. In *Proceedings of the DLM-Forum 2002*, pages 218–227. Office for Official Publications of the European Communities, 2002.

[3] C. A. Lee and K. Woods. Automated redaction of private and personal data in collections: Toward responsible stewardship of digital heritage. In *Proceedings of The Memory of the World in the Digital Age: Digitization and Preservation*, pages 298–313, New York, NY, USA, 2012. UNESCO.

[4] J. Metz. EWF specification – Expert Witness Compression Format specification. https://github.com/libyal/libewf/wiki, 2006.

[5] S. Misra, C. A. Lee, and K. Woods. A Web Service for File-Level Access to Disk Images. http://journal.code4lib.org/articles/9773, 2014.

[7] `http://www.flusser-archive.org/`

[8] HTTP range request support is required to avoid transferring the complete disk image to the emulator's site.

[9] Functional access to the Flusser machine. `http://hdl.handle.net/11270/2b87de90-37dc-4d66-a9e6-546a80b0b261`

Developing a Highly Automated Web Archiving System Based on IIPC Open Source Software

Zhenxin Wu, Jing Xie, Jiying Hu, Zhixiong Zhang
National Science Library, Chinese Academy of Sciences
33 Beisihuan Xilu, Zhongguancun
Beijing P.R.China ,10019
+86-(10)-82628382
wuzx,xiej,hujy,zhangzhx{@mail.las.ac.cn}

ABSTRACT

In this paper, we describe our development of a highly automated web archiving system based on IIPC open source software at the National Science Library (NSL). We designed a web archiving platform which integrates with popular IIPC tools, as well as developing several modules to meet special requirements of the NSL. We have applied a cooperative mode of central management server and collecting client, which can complete the unified management of seeds and support the collaborative work of multiple crawlers. Some modules were developed to improve the automation of web archiving workflows and provide more services.

General Terms

Infrastructure challenges; Frameworks for digital preservation; Preservation workflows; Innovative practice.

Keywords

Open source software, Web archive, Platform development Process automation.

1. INTRODUCTION

Web information, which is considered to have cultural heritage value, is protected under laws in many countries. Web archiving refers to the activities of capturing, preserving and delivering web information over time. It provides a reliable way to preserve the web information permanently and effectively. Far more than one hundred projects are ongoing all over the world.

In science and technology (S&T) fields, a large amount of information is published on the Web. The emphasis of international web archiving activities has steadily been shifted to S&T information on the Internet. The National Digital Information Infrastructure Preservation Program (NDIIPP) published a report called "Science @ Risk: Toward a National Strategy for Preserving Online Science" [1], which shows that preserving online science has explicitly become a national strategy.

The important web information of S&T has become an indispensable part of open resources. With keen awareness of the significance of web archiving, the National Science Library (NSL), Chinese Academy of Sciences has paid close attention to web archiving practices since 2006, and carried out research with funding support from Chinese National Social Sciences. In 2013, NSL began to develop a platform for archiving the important web information of S&T. In this paper, we describe our practice of developing a highly automated web archiving system (NSL-WebArchive) based on IIPC open source software. A highly automated platform, which greatly reduces manual work, offers an important advantage for web archiving in the long term.

2. EXTENSION OF WEB ARCHIVING FRAMEWORK BASED ON IIPC OPEN SOURCE SOFTWARE

2.1 Basic Web Archiving Framework of IIPC

The International Internet Preservation Consortium (IIPC),[1] which was founded in 2003, has more than 40 members from over 25 countries, including national, regional and university libraries and archives and non-profit organizations and commercial service providers. It promotes international cooperation and resource sharing.

IIPC has funded a variety of web archiving tools that can be used to select, harvest and archive Web information, like Heritrix[2], Web Curator Tool (WCT)[3] Wayback[4], NutchWAX[5]. And these tools have been widely applied around the world. The most popular four tools cover basic web archiving, as well as WARC [2], which has been international standard web archive format (ISO 28500).

Only a few web archiving projects have been launched in China, and there is a lack of cases of utilizing the above-mentioned open source tools to design a large-scale web archiving system. So far, the National Library of China is the only institute in China to have deployed the experimental system based on the IIPC framework and has carried out archiving activities for several years.

2.2 Specific Needs of the NSL

According to practices as reported in the literature, the web archiving framework of IIPC often needs to be enhanced or adapted to meet local needs. On the one hand, NSL-WebArchive will harvest large-scale web information periodically, and on the other hand the harvest frequency and the harvesting speed should be low enough so that it will not affect daily access. This causes a

[1] http://netpreserve.org/

[2] https://Webarchive.jira.com/wiki/display/Heritrix/Heritrix

[3] http://Webcurator.sourceforge.net/

[4] http://sourceforge.net/projects/archive-access/files/wayback/

[5] http://archive-access.sourceforge.net/projects/nutch/

tension between harvest cycle and harvesting speed. Meanwhile, the more crawling tasks, the more manual labor will be involved, so automation of large-scale, distributed Web information harvesting and in-depth analysis of archived information, became the key issues to be resolved when developing NSL-WebArchive. At the same time, there is a need to support in-depth analysis services of archived information.

(1) Develop NSL Local Web Archive Management Tools

IIPC has funded a variety of web archiving tools for managing the web harvesting process such as Netarchive Suite and the WCT. But they do not meet our requirements for several reasons.

First, NSL-WebArchive provides access and analysis services based on subjects. We add more descriptive information for the target sites, including institution type, subject area, important research fields, etc. We can provide content-based faceted search, site browse and personalized recommendations. Second, in order to achieve crawling efficiently, we need to get more information about the process of crawling to adjust collection strategies. Third, to develop a highly automated web archiving system, we need to monitor and manage the process of crawling, including the running status of multiple crawlers and the sites that are being crawled. If we use the open source software, we must spend a lot of time analyzing source code and developing additional functions.

Considering the pros and cons, we decided to reuse a product of another project undertaken by our team to develop web archive management platform. Moreover, the National library of France and the British Library have both developed a scheduling management platform to achieve better management results. The British Library has visited our institution for in-depth communication. During the development process, we have given serious consideration to their experiences and lessons.

(2) Enhance Distributed Heritrix Framework

The project is currently in its initial stage. In order to save funds, the computers are not powerful and the configuration is at a low level. The number of sites crawled by Heritrix in parallel on a single server is limited. To improve collection efficiency, we develop a distributed Heritrix Framework, so a number of sites can be crawled at the same time. This framework has two advantages:

A) A number of sites can be crawled in parallel at the same time. So one site can be crawled slowly enough to reduce the pressure for both the crawler computer and remote web site server.

B) One crawling task can be dispatched to different crawler computers randomly, so crawling behavior can be marked as different IPs, and will be likely to be regarded as attack behavior.

(3) Enrich Full-Text Retrieval Function

We use solr cloud as a full-text search engine, so the platform can provide not only full-text retrieval but also faceted retrieval and facet navigation. These functions can support the data analysis module in our future work.

2.3 Extension of Web Archive Framework

Based on the IIPC framework, the NSL has designed an extended solution. See Figure 1 below (particularly the parts with blue lines).

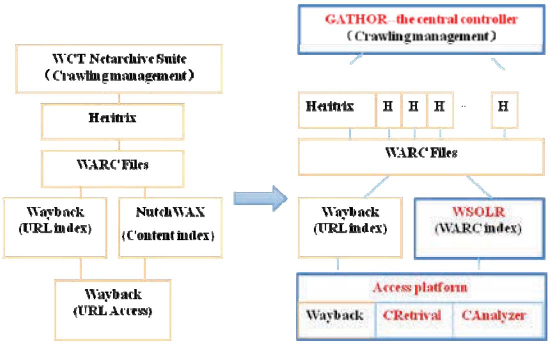


Figure 1. The extended web archive framework

2.3.1 Implementing efficient distributed web archiving management

NSL-WebArchive intends to crawl web information of a relatively fixed and clear website group and does an entire domain crawl for each seed. As the most popular crawler, Heritrix is the best choice for NSL-WebArchive.

Because of so many seeds and internet etiquette, NSL-WebArchive has to deploy many crawlers to execute distributed harvesting tasks at low frequency and speed. The number of crawlers can be increased or decreased according to the tasks.

An efficient distributed web archiving management platform is certainly necessary for NSL-WebArchive, which can manage harvesting tasks and control the distributed crawlers to implement crawling.

2.3.2 Developing an easily recognizable naming convention for WARC files

Each instance of Heritrix uses the default naming rules if it is not changed purposely. But if there are multiple Heritrix systems deployed at the same time, the default naming rules of the crawling configuration files and harvesting files of each Heritrix need to be modified, to allow managers to identify and manage WARC files easily and effectively.

So, an easily recognizable WARC file naming convention becomes necessary. When designing the naming rules, we have had to take many things into account, such as distinguishing these WARC files from different crawlers which are deployed in different servers, and the same seed needing to be collected many times.

2.3.3 Implementing highly-automated processes

Due to a larger number of crawling tasks that need to be configured, managed and periodically scheduled as well as quality control of crawling, we need to realize the automation of crawling task management to reduce manual work.

Multiple distributed crawlers have been deployed in NSL-WebArchive. Unfortunately, Heritrix cannot store WARC files in a remote server, but only in a specified directory of a local server. Each Heritrix has its own result directory even if they are in the same server. Additionally, Wayback can only provide automatic indexing and browse or access service for a specified local directory, so one Wayback cannot work for different Heritrix systems at the same time. NSL-WebArchive will provide a solution for collecting the WARC files from different crawlers in order to facilitate the subsequent management or service.

Without a Hadoop system to use NutchWAX, NSL-WebArchive intends to develop an alternative WARC full-text indexing tool-WSolr (WARC Solr)

2.3.4 Enrich the ways to use archived information

Users need more ways to use archived information. Based on the above-mentioned Solr index, NSL-WebArchive adds a retrieval module named CRetrival，which can provide full-text retrieval and faceted browsing according to subject, timestamp and site, etc.

Finally, NSL-WebArchive intends to supoort content mining and analysis by developing the CAnalyzer module in the future.

3. BUILD UP NSL-WebArchive Platform

Based on the above requirements, we have designed the platform framework with the following three basic principles:

1) The platform framework will integrate with open source software and the customized modules which are developed by the NSL, so that the platform can make full use of the advantages of open source software as well as meet local requirements. And this platform can be built in a short time with better compatibility and seamless upgrade.

2) A cooperative model of central management server and collecting client is applied, which can complete the unified management of seeds and support multiple crawlers' collaborative work.

3) Some modules are developed to improve the automation of web archiving workflows and provide more services.

3.1 NSL-WebArchive Function Framework

NSL-WebArchive applies a cooperative model of central management server and collecting client so that it can implement a distributed crawling and archiving system. As shown in Figure 2, there are three levels, collection level, storage level and access level.

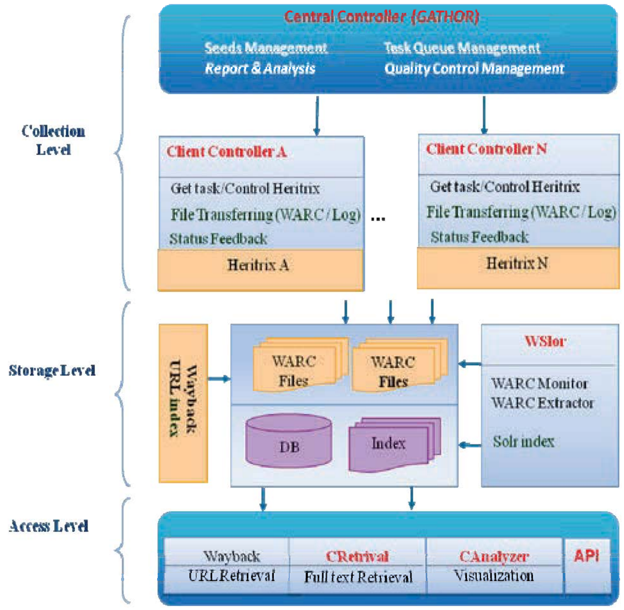


Figure 2. NSL-WebArchive function framework.

3.1.1 Collection Level

The central management server is responsible for the configuarion and management of crawling seeds, and generating and managing the crawling task queue. Meanwhile, the central management server can monitor the status of each crawling task by receiving a report from each client in time.

Each collecting client contains a client controller and an instance of Heritrix. The client controller gets a new crawling task from the task queue of the central management server, and controls Heritrix to crawl web information from the Internet until the crawling task is finished. Then, the WARC files which are stored on local disk of the collecting client will be transmitted to the specified directory in remote server through an FTP pipe, and the current crawling task report will be recorded in database of the central management server.

3.1.2 Storage Level

The storage level stores all WARC files from each collecting client. In addition, we use Wayback and WSolr to create index files in order to provide retrieval and access services.

3.1.3 Access Level

The access level integrates Wayback, CRetrival and CAnalyzer. It provides a series of services, including URL retrieval, content-based retrieval, content analysis and visualization services. APIs will be provided for other system calls, which will be convenient for researchers who are interested in analysis and use of the archived data.

3.2 Workflow of NSL-WebArchive Platform

The workflow supported by the NSL-WebArchive Platform is shown in Figure 3.

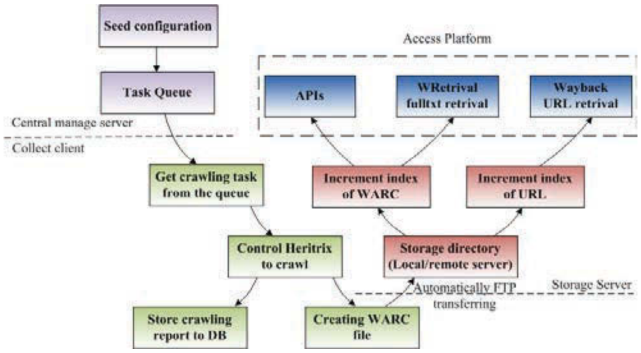

Seed configuration
Access Platform
Task Queue
APIs
WRetrival fulltxt retrival
Wayback URL retrival
Central manage server
Collect client
Get crawling task from the queue
Increment index of WARC
Increment index of URL
Control Heritrix to crawl
Storage directory (Local/remote server)
Automatically FTP transferring
Storage Server
Store crawling report to DB
Creating WARC file

Figure 3. Workflow of NSL-WebArchive platform.

(1) The Manager configures and manages the seeds on the central management server. According to the configuration of each seed, the server will automatically generate the crawling task and put it into the queue on schedule.

(2) The collecting client gets a task from the task queue of the central management server and controls Heritrix to crawl web information from the Internet as well as monitoring the status of Heritrix. When each crawling task is completed, the client will automatically transmit WARC files to the specified directory in the remote server, and then delete the WARC files on its local disk. Finally, crawling logs which are generated by Heritrix for each task will be abstracted and stored in management database of the central management server, and be ready for supporting further analysis and quality control.

(3) Wayback will automatically monitor the specified directory, create an index of the new uploaded WARC files, so users can directly access the new archived data through Wayback.

(4) Similarly, WSolr will automatically monitor the specified directory, extract related information and create incremental Solr index for the new uploaded WARC files, so users can do full text retrieval and facet navigation with CRetrival.

3.3 Advantages of the Collecting Client Active Mode

The NSL-WebArchive Platform is a distributed system, including one central management server and multiple collecting clients. In this system, a definite advantage is the active mode initiated by the collecting client. This platform established an RMI [6] communication pipeline between the central management server and the collecting client. The collecting client actively obtains new crawling task and reports its status to the central management server, so the central management server needs not query each collecting client, and reducing the pressure on both sides, the server and the client. If one collecting client is down, crawling tasks will be assigned to other collecting clients automatically. Unexpected events will not affect the whole platform, and the crawling task will not fail out.

The task token -- which contains the whole description of the crawling task -- is a key element of this distributed system. In one communication between the central management server and the collecting client, the client receives a new task token, decrypts the token, gets crawling task information, and controls Heritrix to carry on the crawling task.

The task token contains: task ID, seed URL, crawling domain, crawling speed and pressure,crawling frequency, seed configuration parameters, etc.

3.4 Developing Multiple Modules to Enhance Process Automation

3.4.1 Task Scheduling Module of the Central Management Server

NSL-WebArchive needs to do a lot of management work, such as seeds management, crawling task configuration, periodically scheduling tasks and quality control. The central management platform implements automated cyclic operation of tasks through a task scheduling mechanism.

This task scheduling mechanism requires the administrator to specify settings for each site collection, including the collection depth, collection frequency, maximum collection time, maximum download, maximum number of jumps, maximum path depth, and the collecting period.

Then the central management server periodically generates collection tasks by setting the timer. The management server periodically checks the collecting period of all sites, and determines whether a new task should be created for a site. If it is overdue, this new task will be put into the job queue.

The collecting client actively obtains tasks from the task queue which is generated by the management server, and generates the necessary configuration file for Heritrix and then calls Heritrix to start collecting. When the task is finished, it obtains the task from the management server again.

In short, once crawling tasks have been configured correctly, the task scheduling module (task scheduler) can dispatch a large number of tasks periodically with the task scheduling mechanism.

3.4.2 WARC File Collecting Module of the Collecting Client

The collecting client periodically executes crawling tasks through the workflow mechanism. The entire process includes some functional modules, from actively obtaining collection command to sending reports of collection results.

Because WARC files are created in different directories by different Heritrix systems, we have developed a collecting module (WARC Gather) for the automatic collection of WARC files. After the collecting client monitors the end of the collection task, WARC Gather transmits WARC files to the specific directory in the remote server by using FTP. After uploading successfully, these local WARC files are deleted. Meanwhile, it transmits log files to the central management server by using the same method.

This module not only solves the remote storage problem of Heritrix, but it also automatically collects WARC files from multiple distributed Heritrix systems.

3.4.3 Status Report Modules

The log files of Heritrix can be uploaded to the management server by WARC Gather. Then the log analysis module of the management server deals with these log files and parses out all sorts of the collection status parameters of each URL and stores them in the database.

Status reports include:

1) The basic report include consumption of time, the number of successful URL, the number of failure URL, the amount of data downloaded, etc.
2) The senior report include proportion of document type, proportion of HTTP status code, seed collection information, URL list and error analysis. All the information is stored in the management database. By adding the task ID to Heritrix source code, statistical data of each crawling task can be viewed.

There is another report status module in the collecting client. By automatically analyzing Heritrix logs, this module monitors the crawling status of Heritrix and presents the crawling status to the central management server whenever necessary, such as the ending of a crawling task or any interruption of a crawling task.

3.5 A Standard Naming Convention

There are four kinds of files that need an effective standard naming convention in NSL-WebArchive.

3.5.1 Seed File

Each crawling task will need a seed file which is created by the client controller after it gets a task from the task queue of the central management server. This seed file is used to store the URL of the target site for Heritrix.

The naming format for seed file is “site domain-seeds.txt”.

3.5.2 Configuration File

Each crawling task will need a configuration file to store crawling parameters for Heritrix.

The naming format of configuration file is "site domain.xml".

[6] http://download.oracle.com/javase/tutorial/rmi/index.html

3.5.3 Task Folder and Task File

Heritrix will generate a task folder for each task in which the crawling log and report are stored. In order to manage the task more easily, we put all the tasks of each month into one sub-folder named with "year-month" in the task folder, e.g., 201403, 201404, 201405 and so on. The task folder "201403" means that it is generated in March 2014 and stores all the tasks that are carried out in that month.

The naming format of task file is "site domain- timestamp"

The UTC time zone is employed for the creation of the time of task folder. Its timestamp format is "yyyyMMddHHmmss".

3.5.4 WARC Storage Directory and WARC File

The WARC files are stored in the remote storage server. The collecting client automatically generates a (new) folder in when it uploads WARC files. As mentioned above, the naming format of each folder is "year-month".The naming format of WARC files is "site domain-WARC file creation time -serial number-Hostname".

1) The site domain is used as prefix to the file name.
2) The WARC file creation time employs UTC time zone. Its format is “yyyyMMddHHmmss”.
3) Serial number is the sequence number of WARC files generated in each crawling task. The WARC file size is predefined.

Take www.las.ac.cn for example:

Task folder is www. las.ac.cn -20140323084011.

WARC file is www.las.ac.cn-20140323084024-00000-Hadoop-master-180.warc.gz

3.6 Developing WSolr and CRetrival

WSolr includes three functions: automatic monitoring of WARC files, content extraction of WARC files, and incremental indexing of Solr.

WSolr uses the same mechanism as Wayback to realize automatic monitoring. Meanwhile, it uses three underlying classes of Wayback, WARCReaderFactory, WARCReader, and WARCRecord to parse the content of WARC files. These modules are used to extract and analysis WARC files.

CRetrival can provide content-based search. It can also provide faceted search of archived sites according to time, subject and resource types. By analyzing crawling logs of Heritrix, it can also provide status summary of each crawling task for each seed.

By extracting data from WARC files, NSL-WebArchive not only enriches the search and access services, but it also lays a good foundation for further services of data mining and data analysis.

The goal of WAnalyer is to do an in-depth analysis of archived content by using visualization techniques. At this moment, it is still in the planning stages. A detailed description of this module is not within the scope of this paper.

4. ANALYSIS OF RUNNING NSL-WebArchive

The NSL-WebArchive platform was complete and put online as a beta version in May 2014. 228 seed sites have been periodically crawled and archived. Until Sept. 2015 a total of 20 TB data (compressed) had been archived. The total number of WARC files is more than **1,200** and the total number of URL is **11,392,701**.

Overall, the NSL-WebArchive platform has achieved good results, which are described as follows.

1) The central manage server provides more effective management functions, which reduced the manual work greatly.
2) By developing multiple modules, NSL-WebArchive significantly improves the degree of automation.
3) WARC file extraction module and Solr faceted indexing not only enriches data retrieval, but also lays a good foundation for the further services of data mining and data analysis.

5. EPILOGUE

The NSL-WebArchive platform not only archive the cultural (science) heritage, but also use data mining to support effective assessment of S&T policy, strategic decisions, trends analysis of domain analysis, and predict future trends, etc.

By developing the NSL-WebArchive platform, NSL has accumulated experiences on large-scale web archiving, especially on system management, scalability, automation, and information reuse. In future work, we need to optimize the crawling strategy by analyzing crawling logs, to enhance data preservation and management of WARC files registration and data backup.

6. REFERENCES

[1] National Digital Information Infrastructure and Preservation Program. 2012. Science @ Risk: Toward a National Strategy for Preserving Online Science. Library of Congress, Washington, DC.

[2] ISO 28500:2009 Information and Documentation -- WARC File Format.

A Method for the Systematic Generation of Audit Logs in a Digital Preservation Environment and Its Experimental Implementation In a Production Ready System

Hao Xu
DICE Center
University of North Carolina at Chapel Hill
Chapel Hill, NC
xuh@email.unc.edu

Jason Coposky
iRODS Consortium
RENCI
Chapel Hill, NC
jasonc@renci.org

Dan Bedard
iRODS Consortium
RENCI
Chapel Hill, NC
danb@renci.org

Jewel H. Ward
SILS
University of North Carolina at Chapel Hill
Chapel Hill, NC
jewel_ward@unc.edu

Terrell Russell
iRODS Consortium
RENCI
Chapel Hill, NC
tgr@renci.org

Arcot Rajasekar
SILS
University of North Carolina at Chapel Hill
Chapel Hill, NC
sekar@renci.org

Reagan Moore
SILS
University of North Carolina at Chapel Hill
Chapel Hill, NC
rwmoore@renci.org

Ben Keller
iRODS Consortium
RENCI
Chapel Hill, NC
kellerb@renci.org

Zoey Greer
iRODS Consortium
RENCI
Chapel Hill, NC
tempoz@renci.org

ABSTRACT

In a digital preservation environment there is a need for a complete auditing of the change of the system state. A complete log ensures that the properties of the objects in the system can be verified. Modern data management systems such as the integrated Rule-Oriented Data System (iRODS) allow administrators to configure complex policies. Pre- or post-operation, these policies can trigger other state changing operations. In this paper, we describe a method that allows us – given a complete list of state changing operations – to generate a complete audit log of the system. We also describe an experimental implementation of the framework. An important advantage of our method is that not only do we build on sound theoretical foundations, but we also validate the methodology in a production ready environment which has undergone substantial quality control. The implementation of our method can be distributed as a turnkey solution that is ready to deploy, which significantly shortens the gap between theoretical development and practical applications.

General Terms

Infrastructure opportunities and challenges

Keywords

audit log, production system, implementation, digital preservation, policies, automated log generation

1. INTRODUCTION

Researchers and practitioners at the Digital Curation Centre (DCC) have defined digital curation as involving "maintaining, preserving and adding value to digital research data throughout its lifecycle" [10]. A data manager begins curation at the time the collection is assembled or acquired. He or she actively manages the collection in order to "mitigate the risk of digital obsolescence" and "to reduce threats to [the data's] long-term research value" [11]. According to DCC researchers and practitioners, auditing is one part of the active curation of a preservation system, and provides a means to ensure stored data has integrity and may be trusted.

When an organization audits a digital repository, two primary standards are used: ISO 14721:2012 [12], the Open Archival Information System (OAIS); and, ISO 16363:2012 [13], the Audit and Certification of Trustworthy Digital Repositories. The former is an ISO standard and reference model that defines an archive as something "consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community". The latter recommendation is based on the OAIS Reference Model [7]. ISO 16363 defines a recommended practice for assessing the trustworthiness of digital repositories. It may be used for all types of digital repositories, regardless of content type, and as the basis for certification of the archive as "trusted" by independent auditors. An important aspect of such auditing activity is to show that the dynamic behavior of the digital repository and the digital curation activities are actually being implemented with regards to the objects in the digital preservation system. In previous work [1], we've shown that a large part of this type of auditing can be checked by inspecting an audit log of state changes. In this paper, we further develop this idea by providing an implementation framework.

Auditing has important implications beyond digital preservation for its own sake. The healthcare and financial services industries, for example, are subject to government privacy and records retention regulations. Administrators of healthcare data need to be able to prove to regulatory agencies that patient records have only been available upon patient consent. Financial records are subject to retention policies, and need to be protected from tampering.

Auditing can also play an important role in industries that are not subject to extensive regulatory requirements, where it can provide insight into illegal hacking activity. Sensitive internal records–HR data and corporate finances–must be protected from unintended access and release. Auditing, with appropriate detection algorithms, can provide administrators with real-time insight into unusual file system activity. In the event that data is compromised, auditing can provide an evidence trail for prosecution, as well as the ability to deconstruct an attack to develop methods to interrupt similar attacks in the future. Data management auditing provides the ability to guarantee regulatory compliance and to safeguard against malicious activity.

In this paper, we propose a implementation framework that allows us to systematically generate a complete audit log of the system, given a complete list of state change operations. We also describe an experimental implementation of the framework and discuss which features enable such implementation. Another innovation in our implementation is that we use the same policy enforcement mechanism for implementing application domain policies to implement auditing, making auditing part of the policies. This reduces duplicate code paths and enable higher test coverage. It also enables interesting use cases such as auditing the auditing mechanism itself, and raises questions concerning how to ensure the termination of such auditing rules.

2. THE METHODOLOGY

In previous work [1], we have shown that a digital repository can be seen as a state transition system and policies related to preservation properties can be described in terms of legal or illegal state transitions. With this process, we tie policy certification to checking the legality of a sequence of state changes. This allows us to implement auditing in a digital preservation environment by providing a complete log of the change of the system state. Existing ad hoc methods do not guarantee the completeness of the audit log. To add to the complexity, modern data management systems such as iRODS allow administrators to configure complex policies to meet the requirements of different application domains. Pre- or post-operation, these policies can trigger other state changing operations. Further, the policies can be nested. These policies are usually executed from a policy language, which are sometimes Turing-complete programming languages, in which case the list of operations performed cannot be easily determined statically. Also, because of the complexity of policies, it would be inefficient to keep track of all commands in the rule language. Rather, we only want to audit the commands that change the system state.

An audit log is considered "complete" for making assertions about preservation properties when we capture all state changes. In a computer system, an important type of state change occurs when a state change operation is applied. The majority of state change operations include: user interaction, time triggered operations, and action trigger operations. The notable exception are state changes by hardware failure, which cannot be entirely addressed on the software level. (This type of state change is often partially addressed by redundancy. When redundancy is set up, we can indirectly capture this kind of state change through certain verification and recovery operations, for example, checking the checksum. The method described in this paper is therefore applicable, albeit indirectly, to this type of state change).

In order to systematically capture this type of state change, we need to find a way to systematically enumerate all state change operations and their applications and keep track of every state change operation.

We can systematically enumerate all state change operations by categorizing them by the different types of effects. For example, a subset of operations supported by the integrated Rule-Oriented Data System (iRODS) [3] is shown in Table 1. The list of database operations, resource operations, and network operations, etc. are fixed, whereas other types of plugins such as microservice are designed for extensibility, therefore no fixed operations are listed.

There is a question of whether the list of operations generated thus far is complete. To show the completeness, we can separate the part of the software that changes the state, or is effectful, from the part of the software that does not change the state, or is effect-free. The effect-free part of the system talks to the effectful part of the system through a well-defined application programming interface (API). The operations in the API map directly to the operations we enumerated. If we capture all API calls across the effectful-effect-free boundary, we capture all state changing operations.

Given a complete list of operations, and a mechanism to capture every call of every operation, we can ensure the completeness of the log. Immediately before and after the application of the operations, we record the event in the log. The implementation, which we will go into details in the next section, will discuss how we capture this information in a production ready system. A complete history of the system can be reconstructed when an administrator inspects the log. By providing the availability of the history, we can verify that the digital repository is compliant with

Plugin Type		Plugin Operation
Resource		create open read write stagetocache synctoarch registered unregistered modified resolve_hierarchy rebalance
Authentication		establish_context agent_auth_verify
Network		client_start client_stop agent_start agent_stop read_header read_body write_header write_body
Database	replica	reg_replica unreg_replica
	data object	reg_data_obj rename_object move_object
	collection	reg_coll_by_admin reg_coll mod_coll rename_coll del_coll_by_admin del_coll
	metadata	mod_data_obj_meta set_avu_metadata add_avu_metadata_wild add_avu_metadata mod_avu_metadata del_avu_metadata copy_avu_metadata del_unused_avus
	resource tree	add_child_resc reg_resc del_child_resc del_resc mod_resc mod_resc_data_paths mod_resc_freespace get_hierarchy_for_resc substitute_resource_hierarchies
	zone	reg_zone mod_zone rename_local_zone del_zone get_local_zone
	user	del_user check_auth make_temp_pw mod_user make_limited_pw reg_user
	access control	mod_access_control gen_query_access_control_setup
	quota	calc_usage_and_quota set_quota check_quota
		start open close rollback commit
Microservice		<microservice_name>
API		<api_name>

Table 1: Plugin Operations

the pre-established policies. Furthermore, the API can be modularized such that the effectful part can be encapsulated into modules and they can be loaded dynamically at run time. This provides flexibility of features yet still guarantees the completeness of the audit log.

To prevent users from inadvertently circumventing our software abstraction,[1] we ensure that the user can only modify the system state through these operations by virtualizing the storage and providing strict access control. The virtualization of the storage ensures that the users are not exposed to low level APIs that could potentially modify the system by bypassing the system-provided operations. Strict access controls ensure that the user cannot inadvertently bypass the virtualization.

3. IMPLEMENTATION

We describe an implementation of our method in iRODS. We choose to implement our method in iRODS because it provides several key features that enable a direct translation of our framework code. Also, the industrial level code quality allows us to bring our implementation to the production system.

iRODS is a state-of-the-art open source software system for addressing the key data management tasks that face users as the size and complexity of digital data collections continue to grow rapidly. Because the principal data management tasks are highly interrelated, rather than taking a piecemeal approach or addressing just a single task, the iRODS system takes a comprehensive approach to full data life-cycle management.

At the same time, the system design is highly user-driven and avoids the pitfalls of a “one size fits all” design by building on a comprehensive generic platform with a highly configurable architecture. In addition, iRODS offers multiple paths to interoperation with outside systems such as repositories, interfaces, and applications. This lets users adapt iRODS to the details of their own environment in a wide range of production applications that can emphasize different aspects of data management in diverse domains.

Furthermore, iRODS has undergone strict quality assurance. We repaired over 1100 identified defects in the 4.1 core code. Using Coverity alone has vastly improved iRODS stability, and coupled with the other tools deployed within our continuous integration (CI) infrastructure, iRODS is in an enterprise production-ready state. In continuous topology testing of multiple machines, our JSON-based Zone descriptions are now ingested by an Ansible-driven engine which deploys a full iRODS topology into our VMWare cloud infrastructure. The current basic test deployment runs a full feature testing suite from multiple types of configurations on every commit to our GitHub repository.

In the nine years since iRODS was first released, the software has been adopted for the support of a variety of research activities. iRODS is in use at over one hundred universities around the world, not only for preservation activities in digital repositories, but also in support of domain-specific research. This utility has begun to spread into the commercial sector, beginning with the life sciences industry. Bioinformaticians use iRODS for its ability to associate data with user-defined metadata and to track the provenance of data as it matures from raw data into a final work product. Gradually, iRODS uptake has begun to spread to other fields, with proofs of concept emerging in the oil exploration and entertainment industries. We expect that iRODS will continue to find use in additional fields, such as the financial services and manufacturing industries.

In the following subsections, we describe iRODS components that enable the design of a high performance auditing system, and an overview of how the auditing system is imple-

[1] Defending against Byzantine error is out of the scope of this paper.

mented.

3.1 Plugin architecture

iRODS has a plugin architecture. This can be seen as a design effort is to move all effectful operations into plugins, and leave the core effect-free. This separation of effectful code from effect-free code allows us to make assertions about state-changing operations through just the observation of interactions with plugins, by defining rules that are dynamically enabled with the dynamic loading of plugin operations. Besides the default supported plugins, the set of supported effectful operations can be extended through microservice plugins.

3.2 Policy enforcement points

iRODS implements the concept of pre- and post-operation policy enforcement points, or PEPs. These PEPs allow system administrators to define rules to be executed either before or after each operation. The policies in a preservation system can then be encoded as rules.

iRODS contains two types of built-in PEPs:

Pre- and post- operation PEPs: these PEPs are triggered before and after an operation is executed. Each operation has a pair of pre- and post- PEPs. User defined rules can be executed at these PEPs to customize the execution of the operations.

Configuration PEPs: these PEPs are triggered at certain points of configuration. Each configuration has one PEP. User defined rules can be executed at these PEPs to customize the configuration of the system.

Built-in PEPs can be extended by dynamic PEPs. For every plugin operation that is called, two policy enforcement points are constructed (both a pre- and post- variety), and if it has been defined in any other loaded rulebase file, they will be executed by the rule engine. The PEP will be constructed of the form `pep_`P`_pre` and `pep_`P`_post`, where P is the operation. For example, for `resource` plugin type, `create` operation type, the two PEPs that are dynamically evaluated are `pep_resource_create_pre` and `pep_resource_create_post`. If either or both have been defined in a loaded rule base, they will be executed as appropriate.

A formal definition of the semantics of PEPs are given in [2]. The flow of information from the pre- PEP to the plugin operation to the post- PEP works as follows: `pep_`P`_pre` should produce information that will be passed to the calling plugin operation. The calling plugin operation will receive any information defined by `pep_`P`_pre` and will pass its own information to `pep_`P`_post`. `pep_`P`_post` will receive any information from the calling plugin operation. A map data structure is made available within the running context of each dynamic PEP based on the plugin type of interest. They are available via the rule engine in the polices.

For example, when running

```
iput -R myOtherResc newfile.txt
```

a `create` operation is called on a resource plugin to create the file. This delegates the call to the actual plugin instance's `create` operation. When `pep_resource_create_pre` PEP rule is evaluated, the values about the file are available for the policy. This allows rule authors to make decisions at a per-resource basis for this type of operation.

3.3 Pluggable rule architecture

The policies are defined at pre- and post-operation PEPs as rules. These rules are executed through a set of rule engines. The pluggable rule architecture allows multiple rule engines to be dynamically and concurrently loaded. Different rule engines can support different languages with the libraries of that language. Every rule engine is automatically equipped with the capability of calling microservices through a single interface. Through the same interface one rule engine can call rules across the rule engine boundary from another rule engine.

This way different rules can be written while taking advantage of the features of different languages, yet still work coherently together. Full compatibility is guaranteed by design with rules written for earlier version of iRODS. Currently, the available rule engine plugins include the iRODS rule language and Python. High performance, natively executed rules can also be written in C++, eliminating the need to go through the microservice interface. Our implementation takes advantage of this capability to provide high performance auditing of the system.

3.4 Auditing policies plugin

The semantic goal of the auditing policy plugin is to provide a complete auditing history to the system without significantly modifying the behavior of the system, including the built-in behavior of the operations and user defined policies. By "not significantly", we mean, low runtime overhead and no change to the semantics of the operations[2].

The auditing plugin provides a turnkey solution to providing the auditing capability to an existing iRODS deployment. The pluggable rule architecture allows users to enable auditing through one switch without interference with existing rules in the system. The code is written in C++ and is compiled and run natively, imposing a much smaller overhead compared to written in an interpreted language such as Python or the iRODS rule language. The events can be arbitrarily filtered, further reducing the overhead for diagnosing a specific type of issue.

The auditing plugin is implemented as a rule engine plugin. The plugin listens to a specific set of events on the server. This set of events include all plugin operation calls. When it receives the event of a plugin operation call, it serializes the calls and the parameters and writes them to the log. This way the log can be parsed and sent to the ELK stack for analysis (Elasticsearch, Logstash, and Kibana) [5].

Since we have shown that the audit log is complete with regard to effectful operations, we can ask the same question of the auditing mechanism itself. How do we know that the

[2] in contrast to data management policies which may change the semantics of operations

audit mechanism does what it says it does? How do we audit the auditing rule? Why not let the auditing rules audit their own execution, which would close the loop?

Letting the auditing rules audit themselves may lead to an infinite loop. Consider the following example: User A initiates Action B. Prior to Action B, the auditing rules are triggered. This leads to an action which is the execution of the auditing rules, before which the auditing rules are triggered again to audit this new action. This leads to an infinite loop. This rules out the simple solution of self-auditing, if we want the auditing rules to terminate.

This is analogous to Russell's Paradox [4], and a classic solution is stratification – we can define a hierarchy of rule execution levels, the lowest being normal rule execution. Each upper level is responsible for auditing the level below. This way, we can provide arbitrary levels of auditing. However, this approach has the following limitation: the execution of the highest level auditing rules is not audited by any other level. We have to trust that they do what they say they do. This can usually be remedied by extensive testing.

More formally, we assign an integer "level" to each action. The normal actions are on level 0. The action of execution of auditing rules triggered by level x action is on level $x+1$. We define a cutoff level, say 2, such that actions of this level do not trigger auditing rules. This allows us to show that the rules for generating the audit log always terminate, which is necessary, because a diverging policy modifies the underlying system in a significant way.

4. RELATED WORK

Currently, practitioners and researchers in the digital library community have developed a series of self-auditing mechanisms and independent certification of a repository as "trustworthy". The Center for Research Libraries [6] has audited a handful of digital libraries and archives and certified them for trustworthiness based on ISO 16363:2011 and ISO 14721:2012. Further work is ongoing to define the requirements for certification of an organization that wishes to provide certification services [8, 9], and to define how those certification requirements will be upheld and monitored themselves. The research outlined in this paper provides a method for proving that required state changes have occurred when certifying a digital repository against a set of policies.

5. CONCLUSION

In this paper, we propose a framework that allows us to generate a complete auditing log of the system, given a complete list of state changing operations. We also describe an experimental implementation of the framework in iRODS.

6. ACKNOWLEDGEMENT

This research is partially supported by the National Science Foundation under Grant Number OCI 0940841

7. REFERENCES

[1] Ward, Jewel H., et al. "Using Metadata to Facilitate Understanding and Certification of Assertions about the Preservation Properties of a Preservation System." Metadata and Semantics Research (2013): 87.

[2] Xu, Hao, et al. "Building an extensible file system via policy-based data management." Proceedings of the 1st ACM International Workshop on Programmable file systems. ACM, 2014.

[3] Ward, Jewel, et al. The Integrated Rule-Oriented Data System (iRODS 3.0) Micro-Service Workbook. DICE, Data Intensive Cyberinfrastructure Foundation, 2011.

[4] Wikipedia. Russell's Paradox. http://en.wikipedia.org/wiki/Russell%27s_paradox

[5] ELK. https://www.elastic.co

[6] Center for Research Libraries. (2015). Certification and Assessment of Digital Repositories. Retrieved April 19, 2015, from http://www.crl.edu/archiving-preservation/digital-archives/certification-assessment

[7] CCSDS. (2012). Reference Model for an Open Archival Information System (OAIS) (CCSDS 650.0-M-2). Magenta Book, June 2012. Washington, DC: National Aeronautics and Space Administration (NASA).

[8] CCSDS. (2011). Requirements for bodies providing audit and certification of candidate trustworthy digital repositories recommended practice (CCSDS 652.1-M-1). Magenta Book, November 2011. Washington, DC: National Aeronautics and Space Administration (NASA).

[9] CCSDS. (2011). Audit and certification of trustworthy digital repositories recommended practice (CCSDS 652.0-M-1). Magenta Book, September 2011. Washington, DC: National Aeronautics and Space Administration (NASA).

[10] Digital Curation Centre. (2010). What is digital curation? Retrieved April 10, 2015, from http://www.dcc.ac.uk/digital-curation/what-digital-curation

[11] Digital Preservation. (2009). Introduction - definitions and concepts. Digital Preservation Coalition. Retrieved April 10, 2015, from http://dpconline.org/advice/preservationhandbook/introduction/definitions-and-concepts

[12] ISO/IEC 14721. (2012). Space data and information transfer systems – Open archival information system – Reference model. Retrieved April 10, 2015 from http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=57284

[13] ISO/IEC 16363. (2012). Space data and information transfer systems – Audit and certification of trustworthy digital repositories. Retrieved April 10, 2015, from http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510

[14] Steinhart, G., Dietrich, D., and Green, A. (2009). Establishing trust in a chain of preservation the TRAC checklist applied to a data staging repository (DataStaR). D-Lib Magazine 15(9/10). Retrieved April 10, 2015 from http://www.dlib.org/dlib/september09/steinhart/09steinhart.html

Educational Records of Practice: Preservation and Access Concerns

Elizabeth Yakel
University of Michigan
School of Information
4323 North Quad
+1 (734) 763 - 3569
yakel@umich.edu

Rebecca D. Frank
University of Michigan
School of Information
3429 North Quad
frankrd@umich.edu

Kara Suzuka
University of Michigan
School of Education
1600 School of Education Building
+1 (734) 408-4461
ksuzuka@umich.edu

ABSTRACT

Researchers in information science are placing increased attention on data reuse and on what must be preserved with that data to enable meaningful use by scholars within and across disciplines. Although the focus has been on scientific or quantitative data, this paper expands the discussion to qualitative data – specifically digital video records of practice in the field of education. This is an interesting case because researchers and diverse education professionals are interested in reusing this content, though their needs differ. We focus on three issues that raise challenges for preservation and access: file format, context, and dissemination.

General Terms

Institutional opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows

Keywords

Data reuse; Qualitative data; Educational records of practice; Digital preservation; Video preservation; Data access

1. INTRODUCTION

Researchers in information science are placing increased attention on data reuse and on what must be preserved with data to enable meaningful use by scholars within and across disciplines. Yet, most of that focus has been on scientific or quantitative data. Less emphasis has been placed on qualitative data, and when it has been considered, the focus has been on textual data. This paper expands the discussion, looking at preservation and access challenges posed by image-based qualitative data – specifically digital video records of practice in the field of education.

In education, records of practice are "detailed documentation of teaching and learning…taken directly from teaching and learning, without analysis, which enable (people) to look at practice" [5]. In many cases, these records are videos of classroom instruction and student activities, which may or may not be accompanied by contextual information such as lesson plans and seating charts.

2. LITERATURE

2.1 Qualitative Data Archiving and Reuse

2.1.1 Curation & Archiving

"The growing inter-disciplinary use, complexity and size of video data make it important for research data services to understand and support it" [33, p. 4]. One of the earliest examples of preservation and archiving of qualitative data is from the UK. Data from a 1930s social research project known as 'Mass-Observation,' was placed at the University of Sussex in the 1970s [14]. Even today, there are few archives that preserve and provide access to qualitative research data. One of the best known sites for qualitative data is the UK Data Archive, which traces its roots in collecting and curating qualitative data back to the early 1990's with the QualiData Project at the University of Essex. While qualitative data archives are more formalized in Europe [26], qualitative data in the US is often hidden in personal collections of faculty papers [6] [25].

Corti [13] has identified key issues for data archives dealing with qualitative data: (1) setting priorities for acquisition, (2) procedures and standards for processing data, (3) metadata standards for documentation, (4) access procedures for safeguarding data, (5) format, (6) researchers, and (7) funding. She argues that these issues are not unique to qualitative data archives, but that for qualitative data "there is more groundwork to be done" [13]. Although this quotation from Corti is from 2000, fifteen years later the groundwork for many types of qualitative data is still lacking. All of these issues have implications for preservation and access.

2.1.2 Reuse

In spite of the fact that few disciplines have established archives for qualitative data, multiple fields have demonstrated interest in preserving and reusing this type of data. Researchers in such diverse disciplines as nursing [21], history [6], geography [24], anthropology [25], sociology [26], psychology [2], and education [15] have all expressed interest in reusing qualitative data and have outlined disciplinary challenges for reuse.

Among these studies of qualitative data reuse, the focus has been mainly on text- rather than image-based data [3]. However, the use and reuse of video data is increasing as tools become available [33]. In this paper, we present results from a preliminary investigation into preservation and access issues surrounding qualitative data in education, specifically video records of practice that are often contextualized by diverse forms of documentation.

3. Records of Practice in Education

Records of practice in education consist of a wide variety of materials in a number of analog and digital formats. These records include student-, teacher-, and researcher-generated data. Student-generated data includes class work products, such as homework or in-class assignments and assessments. Teacher-generated data includes lesson plans, curriculum excerpts, blank assignment papers and assignment instructions, as well as posters, slides, or whiteboard images displaying work produced during and for lessons. Researcher-generated records include videos of classroom or learner activities and observation notes. They also comprise analytic supports such as transcripts and seating charts and products of analyses such as annotations and coding. The key characteristic of educational records of practice is that they are “artifacts and documentation drawn directly from teaching or classroom interactions, including video representations of teachers’ work with students in classrooms” [4, p. 12].

While educational records of practice come in many formats, image-based recordings of educational settings have been used for over 50 years [9]. The methods of recording have changed and moved from analog to digital, a transformation that has increased the potential for data sharing and use in educational settings, as well as added curation and preservation challenges. Nevertheless, the authentic and first-hand nature of these video-based records of practice makes them uniquely valuable. Marsh and Mitchell [23] identify two primary benefits of video-based records of practice: 1) they capture the complexity of classroom activities and preserve the activities for future reuse that would not otherwise be possible, and 2) they foster dialog and thought for viewers.

3.1 Collections of Digital Educational Records of Practice

There are approximately a dozen collections of digital educational records of practice in the United States available for limited access and use by researchers and/or education practitioners. Some are part of formal repositories, others are curated by private organizations or the data producer. Repositories include the Inter-university Consortium of Political and Social Research (ICPSR) that houses the Measures of Effective Teaching (MET) data representing a longitudinal study of 3,000 teacher volunteers in six different school districts and the Teaching and Learning Exploratory (TLE) at the University of Michigan School of Education which curates a variety of collections, such as the Grand Rapids Elementary Mathematics Laboratory 2012 (GREML2012) collection that documents an intensive a week-long summer mathematics laboratory. Both ICPSR and TLE present unedited or minimally edited data. The National Board for Professional Teaching Standards hosts a highly curated collection of videos of skilled expert teachers drawn exclusively from the board certification process while the Teaching Channel produces thousands of edited videos to highlight different facets of teaching and learning and a few “uncut” videos for special licensing or customers using their paid platform, “Teaching Channel Teams.” This brief glimpse shows how repositories apply various strategies for selection and curation. They also have different approaches to access as well as *how* they contextualize the video collections.

For the field of education in particular, the capture of digital records of practice of teaching is rooted in a long history of using videos for teacher education as well as a shorter history of using video in research to capture classroom activities for study – including inquiries into teaching practices, cognitive processes, learning trajectories, and socio-environmental interactions.

3.1.1 Educational Use

Video records of practice are used for a variety of educational purposes. Video-based case studies are used in teacher education and professional development to help to establish “professional vision, which consists of socially organized ways of seeing and understanding events that are answerable to the distinctive interests of a particular social group” [16, p. 606]. The use of video data to establish professional vision in education has been well-documented [7] [20] [29]. Video-based cases help pre- and in-service teachers develop capacities necessary for teaching such as noticing and knowledge-based reasoning [7] [31] [19] [23].

3.1.2 Research use

Researchers collect and use video data, but are less likely to share it or reuse data from others due to difficulties in navigating the required processes to share data (e.g., issues with permissions for sharing video data), and also the lack of infrastructure to enable sharing video data (e.g., very large files sizes [33]. However, reuse is emerging as a viable alternative or complement to data collection as more collections of video data become available. In spite of these gains in making educational records of practice available, preservation and access issues persist.

4. Preservation and Access Issues for Educational Records of Practice

Data reuse is easier when data circulate within a community of practice rather than across communities [32]. Researchers who share and reuse data within a particular community benefit from shared understandings of context and disciplinary traditions. “Disciplines’ histories as well as the configuration of their research communities are factors that can impact their capacity to contextualize and document their data and processes appropriately” [11, p. 645]. However, educational records of practice are created and used by many professional and disciplinary communities. This presents a unique challenge. Researchers from education as well as other fields such as psychology and sociology seek to reuse educational records of practice. A broad range of educational practitioners (e.g., classroom teachers, school administrators, teacher educators) are also interested in these records. There are few shared understandings and traditions among these groups.

Of the seven issues Corti raises [13], we focus on three particularly pertinent for educational records of practice: format, metadata standards for documentation, and access procedures for safeguarding data. Carlson and Anderson assert “the obstacles … are less technological than social, ethical, legal, and institutional” [11, p. 636]; we find that the issues with qualitative data intertwine the technical, social, ethical, and institutional factors.

4.1 Format

Qualitative data formats can present unique challenges to long-term preservation and access [12]. We focus on key two issues for educational records of practice. First, the sheer number of different file formats represented in a single collection poses difficulties. Second, the commercial or proprietary nature of some data and data analysis systems – along with their file formats – creates difficulties in assuring long-term preservation.

4.1.1.1 Multiplicity of Formats

Collections of educational records of practice contain data in a multiplicity of formats. For example, researchers and educators using video records have moved from watching recordings of classrooms to interacting with video “embedded in complex multimedia databases and accompanied by a variety of instructional materials” [28, p. 38]. A collection of educational records of practice might include video in one or more formats, textual and still image data, and spreadsheets or other analysis outputs. This has implications for repositories and users. For repositories, formats often have to be transformed into preservation formats; for users, files must be converted into more commonly used formats.

Gracy [17] argues that archiving and preserving digital video presents new challenges unique to this material. Gracy [17] and Harvey [18] cite key factors as format obsolescence, authenticity, scalability, and economic incentives to provide preservation services. The resources required to support preservation and access of video data are more substantial than other types of digital data.

These include server space and maintaining video editing, authoring, and annotation software [33, p. 30].

4.1.2 Proprietary and Custom-made Systems

A second format issue arises from the use of homegrown and commercial systems, which rely on proprietary formats for data analysis and access. This impedes future reuse and preservation as the data are often only renderable with particular software which is difficult or costly for a repository to maintain. In contrast, reusers of statistical data benefit from open formats (e.g. csv). Video records of practice in these highly customized homegrown or commercial systems cannot take advantage of this efficiency.

4.2 Metadata Standards for Documentation and Other Means of Creating Context

We expand Corti's approach to describing qualitative data, and address metadata as well as contextual information, which is necessary to enable reuse of both quantitative and qualitative data. Video records of practice are interesting because they are themselves contextual information about the classroom, but they also require additional context for analysis, "Videos allow teachers to peer vicariously into real classrooms, which is the context within which teaching ultimately takes place" [8].

Scholars have noted differences between big data and small data. Abreu and Acker [1] argue that context is more important for small data as it is difficult to regain when lost. We enlarge Corti's discussion of metadata to include contextual information more broadly. Contextual information is necessary to enable reuse of both quantitative and qualitative data. Educational video records of practice are interesting because they are a context but often require additional context to be analyzed.

Carlson and Anderson (2007) [11] describe qualitative data reuse in their comparative case study of four projects across the qualitative-quantitative spectrum. Regarding qualitative data, they conclude that "the one who collected the data and the one who interpreted them were the same person, and this had implications for the potential to meet data reuse requirements, because many assumptions, procedures, processes, and decisions often remained undocumented tacit knowledge" [11, p. 646]. For qualitative data, and educational records of practice specifically, context takes two forms: context as metadata and context as data.

Metadata preserves context, including the technological context for preservation actions and decisions, and the research context for reuse decisions. There are a number of promising possibilities for capturing and making this information available. For example, many digital video educational records of practice are created using the MP4 video format, which has implicit metadata containing details about the file author, the software used in its creation, and the time and date in which it was created, often structured in XMP format. Along with this, there are several metadata standards for describing digital educational materials, such as the *IEEE Learning Object Metadata Standard* [10] and the Learning Resources Metadata Initiative (LRMI). However, these standards have had limited adoption so far. In addition, there are no agreed upon standards or guidelines among educational records of practice producers for recording information about the files. The information applied by the data producers varies widely and is a major concern. Currently repositories often have to apply a substantial amount of metadata to provide access to the digital video records of practice, to make them discoverable, searchable, and useful. This metadata includes (1) descriptive metadata about the content captured on the video (e.g. information about the district, school, classroom, lesson, and students); (2) technical information about the video (e.g. descriptions of the available audio tracks, camera angles, and synchronized text-based tracks); and (3) specialized tags that map the video files or segments *within* the videos to relevant professional standards, frameworks, or rubrics.

Context is also preserved by associated documentation that accompanies the digital video. The amount of context provided varies, depending on the producer's original purposes and designs. Contexts can range from a transcript of the classroom video (e.g. the dataset, *Towards Dialogue: A Linguistic Ethnographic Study of Classroom Interaction and Change* found in the UK Data Service repository[1]) to abundant documentation including lesson videos (recorded from multiple angles) accompanied by a variety of classroom artifacts and supporting documents, such as video "table of contents," transcripts, student written work, lesson plans, classroom images, seating charts, and tags for the applicable standards and key teaching practices (see the *Grand Rapids Elementary Mathematics Laboratory 2012 Collection* in the TLE).

The amount and kind of contextual documentation available influences the types of reuse that are possible. In most cases, research reuse (as opposed to educational reuse) requires more documentation. Educators often focus on the teacher and the teaching techniques of a particular grade level, or content, and may want access to an assignment. Researchers are more likely to be interested in broader contextual information, such as school demographics (e.g., SES of the district, demographics about race or ethnicity of students). The amount of associated documentation has preservation implications. Diverse documentation increases the number of file formats and the number of files which must be tracked in the archival and dissemination information packages. This creates greater complexity in maintaining relationships between individual files as well as their relationships to the collection as a whole (e.g., maintaining links of work from one student). Finally, since data producers often combine external documentation (e.g. demographic or student test scores from the school district) with the video records of practice, intellectual property issues, discussed in the next section, may be important.

4.3 Access Procedures for Safeguarding Data

Dissemination and access are difficult for digital records of practice in education for two reasons. First, logistics can be complicated. Second, confidentiality and privacy issues abound, particularly since many videos feature minors or teachers whose practices data producers do not want scrutinized or harshly judged.

4.3.1 Logistics of access

Two issues stand out in the logistics of access for video records of practice in education: (1) different repository access environments, and (2) data reusers' preferences about how video is presented.

The preservation and access environments for digital educational records of practice are often different. Access environments almost always require transformation or special processing to create a usable dissemination information package. For example, the TLE uses the Kaltura video platform for disseminating streaming videos. Due to the costs, only highly compressed videos optimized for streaming delivery are stored in the Kaltura Cloud. Source files and large derivative files are stored in less expensive, less accessible offline and online locations for preservation purposes. Since video source files tend to be large, many repositories compress and stream

[1] http://dx.doi.org/10.5255/UKDA-SN-850448

video rather than pay to store large source files in high-capacity access systems or try to deliver them over the internet.

Repositories do not always receive source video files. Large file sizes and limitations of bandwidth, time, and other resources frequently result in decisions to compress source videos files – creating entirely new files – prior to delivering them to a repository. In such instances, the original video metadata can be lost if not carefully preserved prior to the compression process. This can create fidelity and integrity issues for researchers. Video compression can also create quality issues for other types of reuse (e.g. (re-)editing videos for new products).

For data reusers interested in using educational records of practice in teaching, there is demand for videos that support different pedagogies. For example, teacher-educators want to use digital video in class as well as have students view, annotate, and integrate parts of the videos into assignments completed outside of class [27]. This range of uses raises issues about the level of data services provided by the repository and the allowable uses given the confidentiality and privacy issues we address next.

4.3.2 Confidentiality/Privacy

Problems around confidentiality and privacy can be more challenging for video than other types of data. Whyte writes, “Legal and ethical issues affect video data more acutely, although they fall into similar categories as for other media; those associated with gathering data and those with making it available for reuse, the distinction also known as ‘rights in’ and ‘rights out’. In both cases the main issues surround rights and responsibilities to privacy and property” [33, p. 33]. Parry and Mauthner [26], Lin [21], and Cliggett [12] all discuss data management issues associated with qualitative data, such as confidentiality, ownership, and anonymity. Confidentiality issues are common in all types of data reuse. For qualitative data (interview, focus group, video) there are particular issues: (1) anonymizing the data, (2) third party information, and (3) the increasing accuracy of facial recognition software.

Qualitative data is harder to anonymize than statistical data. In statistical data, repositories can more easily identify fields most likely to contain confidential information, and assess whether the aggregation of information could lead to loss of confidentiality. In qualitative data there is no demarcation, the entire text or video requires assessment at a more granular level.

Qualitative data contains information about the study participant, but may also reveal information about others. Third party disclosures raise privacy concerns. For example, a video focusing on a teacher may show students or teacher aides. An interviewee may discuss how to handle particular learning problems in a classroom that reveals the identity of a student.

Privacy and confidentiality require special responses from repositories. For example, curators at ICPSR ask data reusers to sign a confidentiality agreement to use the MET data. Then, the video data is delivered through a web browser requiring secure login and non-video data is delivered in the virtual data enclave (VDE), which allows access to confidential data through a virtual machine. When using the VDE, the researcher accesses and manipulates data on a remote server using his or her own computer. This isolates the data from the researcher’s computer because the researcher cannot download, copy, or remove data from the secure environment. In the ICPSR system, the researcher can run analyses on the virtual server and share relevant analytic files with team members [22].

Data reusers are also affected by privacy and confidentiality concerns. Sometimes contextual information is prevented from being shared. “When archived qualitative data are used for secondary analysis, there should be little doubt that the context that informs the data can never be fully disclosed. Thus, "reality" is in some ways lost for a secondary researcher” [3, p. 17].

Finally, the increasing accuracy in facial recognition software and image-based search is making anonymity and confidentiality more difficult. For example, researchers have found Facebook’s facial recognition software, DeepFace, to be over 97% accurate [30]. Although all the repositories with educational records of practice involving actual video classroom data require registration and a confidentiality agreement, the potential harm of disclosure increases as facial recognition technology develops and spreads.

5. Conclusion

Our investigation into the long term curation and preservation of educational records of practice is just beginning. This paper provides a broad view of the landscape and points to how key issues of file format, context, and access procedures are linked to both preservation and access activities. Our next steps are to examine the dynamics of using the data from the perspectives of data reusers and to probe more deeply into how the preservation issues are being addressed by the different repositories.

6. Acknowledgements

This research is funded by the Institute of Museum and Library Services # LG-06-14-0122-14.

7. References

[1] Abreu, A., & Acker, A. (2013). Context and Collection: A Research Agenda for Small Data. In *iConference 2013 Proceedings* (pp. 549–554). Fort Worth, Texas. http://doi.org/10.9776/13275

[2] Adolph, K. E., Gilmore, R. O., Freeman, C., Sanderson, P., & Millman, D. (2012). Toward Open Behavioral Science. *Psychological Inquiry*, *23*(3), 244–247. http://doi.org/10.1080/1047840X.2012.705133

[3] Andersson, E., & Sørvik, G. O. (2013). Reality Lost? Re-Use of Qualitative Data in Classroom Video Studies. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *14*(3). Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/1941

[4] Bacevich, A. E. (2010). *Building Curriculum for Teacher Education: A Study of Video Records of Practice.* Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/77781

[5] Bass, H., Usiskin, Z., Burrill, G., National Research Council (U.S.), Mathematical Sciences Education Board, & United States National Commission on Mathematics Instruction (Eds.). (2002). *Studying Classroom Teaching as a Medium for Professional Development Proceedings of a U.s.-Japan Workshop*. Washington, DC: National Academy Press. Retrieved from http://www.nap.edu/catalog/10289.html

[6] Blodgett, P. J. (2003). Using Our Faculties: Collecting the Papers of Western Historians at the Huntington Library. *The Western Historical Quarterly*, *34*(4), 491–499. http://doi.org/10.2307/25047347

[7] Blomberg, G., Stürmer, K., & Seidel, T. (2011). How Pre-Service Teachers Observe Teaching on Video: Effects of Viewers’ Teaching Subjects and the Subject of the Video. *Teaching and Teacher Education*, *27*(7), 1131–1140. http://doi.org/10.1016/j.tate.2011.04.008

[8] Brunvand, S. (2010). Best Practices for Producing Video Content for Teacher Education. *Contemporary Issues in Technology and Teacher Education*, *10*(2), 247–256.

[9] Burleigh, J. C., & Peterson, H. W. (1967). Videotapes in Teacher Education. *The Elementary School Journal*, *68*(1), 35–38.

[10] Campbell, L. (2007). Learning Object Metadata (LOM). In S. Ross & M. Day (Eds.), *DCC Digital Curation Manual*. Bath, UK: HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils. Retrieved from http://www.dcc.ac.uk/resource/curation-manual/chapters/learning-object-metadata

[11] Carlson, S., & Anderson, B. (2007). What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use. *Journal of Computer-Mediated Communication*, *12*(2), 635–651. http://doi.org/10.1111/j.1083-6101.2007.00342.x

[12] Cliggett, L. (2013). Qualitative Data Archiving in the Digital Age: Strategies for Data Preservation and Sharing. *The Qualitative Report*, 1–11.

[13] Corti, L. (2000). Progress and Problems of Preserving and Providing Access to Qualitative Data for Social Research—The International Picture of an Emerging Culture. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *1*(3). Retrieved from http://www.qualitative-research.net/index.php/fqs/article/view/1019

[14] Corti, L. (2013). Infrastructures for Qualitative Data Archiving. In *Forschungsinfrastrukturen für die qualitative Sozialforschung* (pp. 35–62). Berlin: Scivero. Retrieved from http://www.germandataforum.org/dl/downloads/forschungsinfrastrukturen_qualitative_sozialforschung.pdf#page=26

[15] Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., … Sherin, B. L. (2010). Conducting Video Research in the Learning Sciences: Guidance on Selection, Analysis, Technology, and Ethics. *Journal of the Learning Sciences*, *19*(1), 3–53. http://doi.org/10.1080/10508400903452884

[16] Goodwin, C. (1994). Professional Vision. *American Anthropologist*, *96*(3), 606–633. http://doi.org/10.1525/aa.1994.96.3.02a00100

[17] Gracy, K. F. (2007). Moving Image Preservation and Cultural Capital. *Library Trends*, *56*(1), 183–197. http://doi.org/10.1353/lib.2007.0050

[18] Harvey, R. (2012). *Preserving Digital Materials* (2. ed). Berlin [u.a.]: de Gruyter Saur.

[19] Koc, Y., Peker, D., & Osmanoglu, A. (2009). Supporting Teacher Professional Development Through Online Video Case Study Discussions: An Assemblage of Preservice and Inservice Teachers and the Case Teacher. *Teaching and Teacher Education*, *25*(8), 1158–1168. http://doi.org/10.1016/j.tate.2009.02.020

[20] Lefstein, A., & Snell, J. (2011). Professional Vision and the Politics of Teacher Learning. *Teaching and Teacher Education*, *27*(3), 505–514. http://doi.org/10.1016/j.tate.2010.10.004

[21] Lin, L.-C. (2009). Data Management and Security in Qualitative Research: *Dimensions of Critical Care Nursing*, *28*(3), 132–137. http://doi.org/10.1097/DCC.0b013e31819aeff6

[22] Lyle, J. (2014). OpenICPSR: OpenICPSR. *Bulletin of the American Society for Information Science and Technology*, *40*(5), 55–56. http://doi.org/10.1002/bult.2014.1720400514

[23] Marsh, B., & Mitchell, N. (2014). The Role of Video in Teacher Professional Development. *Teacher Development*, *18*(3), 403–417. http://doi.org/10.1080/13664530.2014.938106

[24] Moore, F. P. L. (2009). Tales from the Archive: Methodological and Ethical Issues in Historical Geography Research. *Area*. http://doi.org/10.1111/j.1475-4762.2009.00923.x

[25] Parezo, N. J. (1996). The Formation of Anthropological Archival Records. In W. D. Kingery (Ed.), *Learning from Things: Method and Theory of Material Culture Studies* (pp. 145–174). Washington, D.C: Smithsonian Institution Press.

[26] Parry, O., & Mauthner, N. S. (2004). Whose Data are They Anyway?: Practical, Legal and Ethical Issues in Archiving Qualitative Research Data. *Sociology*, *38*(1), 139–152. http://doi.org/10.1177/0038038504039366

[27] Rasmussen, Karsten B. (Ed.). (2010). Qualitative and Qualitative Longitudinal Resources in Europe: Mapping the Field and Exploring Strategies for Development [Special issue]. *IASSIST Quarterly*, 34(3&4).

[28] Santagata, R. (2009). Designing Video-Based Professional Development for Mathematics Teachers in Low-Performing Schools. *Journal of Teacher Education*, *60*(1), 38–51. http://doi.org/10.1177/0022487108328485

[29] Sherin, M., & van Es, E. (2009). Effects of Video Club Participation on Teachers' Professional Vision. *Journal of Teacher Education*, *60*(1), 20–37. http://doi.org/10.1177/0022487108328155

[30] Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. In *Proceedngs of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1701–1708). Columbus, OH: IEEE. http://doi.org/10.1109/CVPR.2014.220

[31] Van Es, E., & Sherin, M. (2002). Learning to Notice: Scaffolding New Teachers' Interpretations of Classroom Interactions. *Journal of Technology and Teacher Education*, *10*(4), 571–596.

[32] Van House, N. A., Butler, M. H., & Schiff, L. R. (1998). Cooperative Knowledge Work and Practices of Trust: Sharing Environmental Planning Data Sets. In *Proceedings of the 1998 ACM Conference On Computer Supported Cooperative Work* (pp. 335–343). Seattle, Washington: ACM. http://doi.org/10.1145/289444.289508

[33] Whyte, A. (2009). *Roles and Reusability of Video Data in Social Studies of Interaction. SCARP Case Study No. 5*. Edinburgh: Digital Curation Centre. Retrieved from https://www.era.lib.ed.ac.uk/handle/1842/3380

Poster Summaries

Dash Curation Service Infrastructure Enhancement: An Informed Extension & Redesign

Nancy J. Hoebelheinrich
California Digital Library
University of California Office of the President
415 20th Street, 4th Floor
Oakland, CA 94612-2901
+1-510-987-6482
nancy.hoebelheinrich@ucop.edu

Stephen Abrams
California Digital Library
University of California Office of the President
415 20th Street, 4th Floor
Oakland, CA 94612-2901
+1-510-987-6482
stephan.abrams@ucop.edu

ABSTRACT

University libraries and data repositories are increasingly being asked to support research data curation as a consequence of funder mandates, pre-publication requirements, institutional policies, and evolving norms of scholarly practice. While free commercial alternatives such as figshare and Dropbox provide high service functionality and intuitive user experience that serve research data creators well, they do not offer long term preservation reliability, nor do they necessarily share the increasingly important value of open data. From the perspective of the research data creator, however, all of these factors are important and desirable, so a preservation repository service targeting the needs of researchers should provide them. The UC Curation Center (UC3) at the California Digital Library created its *Dash* research data portal to address these needs. Following the initial deployment of the *Dash* service UC3 received feedback from users that additional functionality and a redesigned user interface would be desirable. With funding from the Alfred P. Sloan Foundation UC3 has re-factored the infrastructure behind *Dash*, and improved the front-end user experience of the existing deposit service. The *Dash* submission, harvesting, and discovery components are being extended to apply to any standards-compliant repository supporting the SWORD submission and OAI-PMH metadata harvesting protocols.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows;

Keywords

Data curation; Data repository micro-services, SWORD submission protocol, OAI-PMH metadata harvesting protocol

1. INTRODUCTION

University libraries and data repositories are increasingly being asked to support research data curation in response to funder mandates, publication requirements, institutional policies, and evolving norms of scholarly practice. While free commercial alternatives such as figshare and Dropbox provide high service functionality and intuitive user experiences, they do not offer long-term preservation reliability, nor do they necessarily share the increasingly important value of open data. From the perspective of the research data creator, however, all of these factors are important and desirable, so a preservation repository-service targeting the needs of researchers should provide them. The UC Curation Center (UC3) at the California Digital Library created its *Dash* research data portal to address these concerns. *Dash* is not a repository itself, but rather a submission and discovery overlay layer sitting on top of CDL's Merritt curation repository that provides drag-n-drop upload, metadata entry, DOI assignment, and faceted search/browse.

2. ENHANCEMENT PLAN

After several years of providing the *Dash* service, UC3 received feedback from users that additional functionality and a redesigned user interface would be desirable, so a proposal was made to and funded by the Alfred P. Sloan Foundation to re-factor the infrastructure behind *Dash*, and to improve the front-end user experience of the existing self-service deposit workflow. The *Dash* service upgrade will continue to use the underlying micro-services architecture of extending repository functionality by developing small, independent, protocol-linked components rather than by adding to large, monolithic systems. Thus, the *Dash* submission, harvesting, and discovery overlay layer is being extended to apply to any standards-compliant repository supporting the SWORD submission and OAI-PMH metadata harvesting protocols. For a more complete picture of the components, and their interactions, see Figure 1, Dash Functional Architecture.

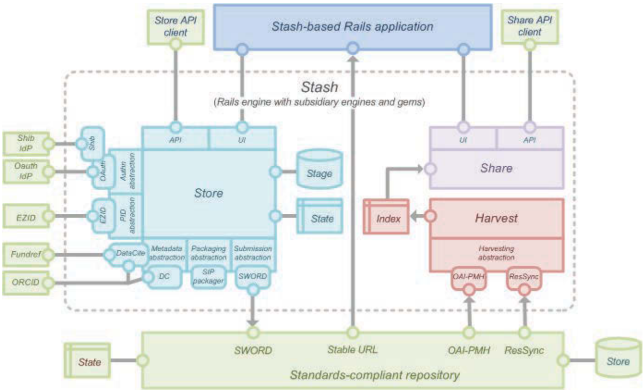


Figure 1. Dash Functional Architecture

3. PROTOCOL SUPPORT

Protocol support is provided by pluggable modules conforming to the APIs of internal abstraction layers for authentication, metadata entry and serialization, persistent identifiers including DOIs and ARKs, repository packaging and submission, and metadata harvesting. Besides supporting the SWORD and OAI-PMH protocols, the *Dash* service will support Shibboleth and OAuth authentication, DataCite and Dublin Core metadata schemes, and EZID metadata management. The front-end user experience is being informed by a more complete suite of user stories in order to provide a simpler, more intuitive interface designed with the individual researcher in mind. Researchers will be able to document, preserve, and publicly share their own data with minimal support required from repository staff, and be able to find, retrieve, and reuse data made available by others.

4. UI/UX-INFORMED DESIGN

One of the reasons often given to explain why researchers do not use repository tools for data submission is the poor design of their user interfaces. Often, the user interface does not take into account the user's experience (or inexperience) and expectations. Because so much of researchers' activities are conducted on the Internet, they are exposed to many high-quality, commercial-grade user interfaces in the course of a workday. Correspondingly, researchers have high expectations for clean, simple interfaces that can be learned quickly, with minimal need for contacting repository administrators. By means of extensive research into the user experience and usability testing, *Dash* is being designed with a simple, intuitive interface that will allow researchers to document, preserve, and publicly share their own data with minimal support required from repository staff, and also be able to find, retrieve, and reuse data made available by others.

5. SEEKING COLLABORATORS

By describing the *Dash* enhancement work in progress, CDL UC3's innovative and generalizable approach will show how the proven *Dash* research data portal, targeted to the needs of individual researchers, is being extended. Besides expanding awareness of the *Dash* service, CDL UC3 staff would like to identify potential collaborators from the digital preservation / open source communities who would be interested in participating and further developing the *Dash* software. More information about the current *Dash* service can be found at: http://dash.cdlib.org and about the *Dash* enhancement project at: https://confluence.ucop.edu/display/Stash/Stash+Home.

An Institutional Digital Repository Backbone

Adi Alter
Ex Libris
Bldg. 8-9 Malcha Technological Park
Jerusalem, 91481
972 2 649 9320
Adi.Alter@exlibrisgroup.com

Ido Peled
Ex Libris
350 E Touhy Avenue, Suite 150 W
Des Plaines, IL 60018
1 617 332-8800
Ido.Peled@exlibrisgroup.com

ABSTRACT

In this poster we will describe how Ex Libris Rosetta serves as a digital repository catering institutions' different needs.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice.

Keywords

DAM; Preservation; Ex Libris; Rosetta.

1. INTRODUCTION

At a time when the amount of digital data produced is growing at a faster pace than ever before, more and more institutions face challenges and are struggling to fulfill the mandates given to them to manage and preserve the digital content being generated in different departments within the institution —libraries, archives, the Institutional Repository, cultural heritage centers, research groups and more. Each of these departments may have different needs, expectations, workflows, policies, data types, requirements for integrations with third-party software, and so on.

Digital content managers need to provide a unified solution that not only meets the present-day requirements of each department, but also offers a high degree of flexibility to support the ever-changing future needs of different users.

2. POSTER CONTENT

In this poster we will present the challenges being faced by different institutions and will illustrate the ways a single digital preservation and DAM solution, used by a wide variety of types of institutions worldwide, can support diverse digital management and preservation activities.

For instance we will display an end-to-end workflow used by an archivist, who would like the system do be integrated with a records management system, will use it as a dark archive and would have retention policies associated with some of the data. We will compare this workflow to a workflow used by a librarian, who would like to the system to be integrated with an Integrated Library System (ILS) and would like the content to be accessible through the institution discovery solution based on various access rights policies. Both would like to use advanced preservation and content migration strategies to make sure their different types of content would be easily accessible also in the future.

Displaying these (and other) workflows side by side in a graphical way will clarify on one hand the different needs that each of these workflows has, and on the other hand will amplify the ways the system handles these needs.

Managing and Preserving Research Data in Ex Libris Rosetta

Adi Alter
Ex Libris
Bldg. 8-9 Malcha Technological Park
Jerusalem, 91481
972 2 649 9320
Adi.Alter@exlibrisgroup.com

Ido Peled
Ex Libris
350 E Touhy Avenue, Suite 150 W
Des Plaines, IL 60018
1 617 332-8800
Ido.Peled@exlibrisgroup.com

ABSTRACT

In this demonstration we will show how Ex Libris Rosetta addresses the challenges of research data management and curation.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice.

Keywords

DAM; Preservation; Research Data; Research Data Management; Ex Libris; Rosetta.

1. INTRODUCTION

Universities now require the life-cycle management of research-related institutional outputs dealing with data from creation to dissemination in a way that ensures compliance with regulations and funding bodies' requirements for the management, protection and sharing of research data.

Many libraries have taken the initiative to help curate research data but lack supporting solutions. In this demonstration we will show how Ex Libris Rosetta provides end-to-end workflows for data curation and archiving of research-related content produced by the institution's research community.

2. DEMO CONTENT

We will discuss the challenges of research data management and curation and will show how Rosetta's innovative open and extendable solution addresses these challenges and is capable of handling any format type produced by researchers, including propriety, diverse and sometimes unrecognized formats, unstructured metadata handling and more.

We will also demonstrate how Rosetta is a research data curation system, with features that deal – for example – with ensuring data integrity, controlling accessibility, recording the impact and traceability of published content, taking into account privacy enforcement, and more. On top of this we will demonstrate preservation strategies in action and show how risks are identified and migration processes are executed on content in risk.

In addition, we will show how Rosetta is dealing with some of the biggest challenges of research data management, like handling gigantic files and huge number of files in a single deposit, addressing the needs of diverse and ad-hoc access rights and policies, handling multiple metadata standards and file formats, and much more.

In the Thicket of It with the NDSA Standards and Practices Working Group: Cultivating Grass Roots Approaches to Real-World Digital Preservation Issues

Winston Atkins
Duke University Libraries
Campus Box 90189
Durham, NC 27708-0189
+1-919-660-5843
winston.atkins@duke.edu

Erin Engle
Library of Congress
101 Independence Ave, S.E.
Washington, DC 20540
+1-202-707-1120
eengle@loc.gov

Andrea Goethals
Harvard Library
90 Mt. Auburn St.
Cambridge, MA 02138
+1-617-495-3724
andrea_goethals@harvard.edu

Karl J. Jackson
United States Marine Band (retired)
8th & I Sts., S.E.
Washington, DC 20390
+1-202-433-4398
karl.j.jackson@gmail.com

Carol Kussmann
University of Minnesota Libraries
499 Wilson Library
309 19th Avenue South,
Minneapolis, MN 55455
+1-612-626-0099
kussmann@umn.edu

Kate Murray
Library of Congress
101 Independence Ave, S.E.
Washington, DC 20540
+1-202-707-4894
kmur@loc.gov

Michelle Paolillo
Cornell University
218 Olin Library
Ithaca, NY 14853
+1-607-255-1038
map6@cornell.edu

Mariella Soprano
Caltech Archives & Special Collections
Mail Code 015A-74
Pasadena CA 91125
+1-626-395-2501
mariella@caltech.edu

ABSTRACT

The engaged membership of the National Digital Stewardship Alliance's Standards and Practices (S&P) Working Group are active digital preservation practitioners. One of five National Digital Stewardship Alliance (NDSA) working groups, S&P projects and discussions originate from real-world issues that members face in their daily work. Since 2010, the S&P has sought to identify community knowledge gaps for the "on-the-ground practitioners" across a broad spectrum of content areas and to work collaboratively to bridge those gaps. Some of the topics recently addressed by the S&P include preservation of digital artworks, issues related to optical media, stumbling blocks for preserving video collections and analyzing risks and benefits of the PDF/A3 format for archival institutions among many others. Using the visual imagery of a fruit tree, this poster explores the grass roots nature of S&P projects and products, from the foundational member institutions comprising the soil and roots, through the trunk and branches of the tree addressing different topics, and finally reaching to the individual leaves and fruit representing project outcomes and deliverables, as well as work still to do. The goal of the poster is to highlight the self-organizing nature of the S&P's varied projects as well as to increase community awareness of the collaboratively developed resources and products.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice; Training and education.

Keywords

Best Practices, Community, Collaboration, Education, Standards, Survey

1. INTRODUCTION

Since 2010, the membership of the NDSA S&P Working Group have come together to discuss current and pressing issues in preserving digital material amongst a set of engaged and active practitioners working in the field. [1] S&P projects and discussions are generated by real-world issues and concerns that members face in their daily work. Working as a community of peers, S&P members have sought to identify community knowledge gaps for the "on-the-ground practitioners" across a broad spectrum of content areas and to work together collaboratively to bridge those gaps. The goal of the poster is to highlight the self-organizing nature of the working group's varied projects as well as to increase community awareness of the collaboratively developed resources and products.

2. THE NDSA

The NDSA is a consortium of institutions that are committed to the long-term preservation of digital information. [2] The NDSA's mission is to establish, maintain, and advance the capacity to preserve digital resources for the benefit of present and future generations. The NDSA comprises over 160 participating institutional members from 45 states and include universities, consortia, professional societies, commercial businesses, professional associations, and government agencies at the federal, state, and local level. The NDSA was launched as a membership organization in July 2010 as an initiative of the Library of Congress's National Digital Information Infrastructure and Preservation Program. [3]

2.1. Standards and Practices Working Group

The S&P melds the expertise from the digital preservation community with practitioners' everyday needs to facilitate a broad understanding of the role and benefit of standards in digital preservation and how to use them effectively to ensure durable and usable collections. The S&P also works actively, in collaboration with other individuals and organizations where appropriate, to identify, promote, and widely disseminate practices found to be effective for selecting, organizing, describing, managing, preserving and serving digital content. The activities and the outputs of the group are shared broadly including the Library of Congress *Signal* blog and the Library of Congress Digital Preservation web site. [4] Since 2011, when the blog launched, S&P-related blog posts, web pages and reports have received more than 16,370 page views.

3. PROJECTS AND PRODUCTS

Work in the S&P originates from the interests and issues of the active membership. Projects germinate organically, often from the groups' round robin "what are you working on?" discussions. An issue is identified from real-world concerns and an action team of volunteers comes together to work on the problem collaboratively. Our range of projects is wide-reaching and varied, reflecting the diverse interests of the membership.

3.1. Media Projects

Optical media: The S&P invited speakers with a varied range of experience to explore issues with optical media. Represented institutions include Library of Congress, AVPreserve, George Blood Audio Video Film, WNYC and BMS/Chace. Topics included data extraction, physical condition issues, and emerging formats.

Software-based Art: The S&P invited experts from four collecting institutions (SFMOMA, MoMA, Rose Golden Archive of New Media Art, and Smithsonian Institution Time Based Media Art project) to share their experiences in both preserving and providing access to digital art works and other new media.

3.2. Format Projects

Video: Many practitioners consider digital video preservation problematic. The S&P hosted several video-related efforts including the Federal Agencies Digitization Guidelines Initiatives (FADGI) reports on comparing file formats for video reformatting and *Creating and Archiving Born Digital Video* as well as video preservation efforts at Stanford, Harvard and NYU Libraries. These led to a "Video Deep Dive" subgroup which developed and conducted the Stumbling Blocks to Preserving Video Survey to identify and rank issues that may hinder digital video preservation.

PDF/A3: The S&P wrote a report that takes a measured look at the costs and benefits of the use of the PDF/A-3 format, especially as it effects content arriving in collecting institutions. [5]

Email: The S&P helped initiate an Email Interest Group to discuss issues, projects and workflows to preserve email; contributed to an Archiving Email Symposium and workshop; and held online tool demonstration sessions with presentations from Harvard, Stanford, Smithsonian Institution Archives and others.

3.3. Content Packaging and Metadata

The S&P organized several sessions around content packaging and metadata. Speakers from the Library of Congress, Harvard, NARA, Georgetown University Libraries and Portico covered packaging forms, SIP components and metadata concepts covered by AS-07 MXF and METS. Other sessions covered tools, practices and workflows for metadata in audiovisual collections. Speakers included AVPreserve (AVCC and Catalyst tools), WGBH, and the Canadian Museum for Human Rights.

3.4. Organizational Practices

Staffing Survey: The S&P conducted a survey to determine how institutions staffed and organized preservation functions, produced an award-winning poster at iPRES2012 along with a detailed report and deposited the raw data in ICPSR.

National Agenda: The S&P contributed significant input and informed actionable recommendations to the *Organization Policies and Practice* chapter of the NDSA 2015 National Agenda for Digital Stewardship. [6]

3.5. Communicating Standards and Practices

Content Integrity/Fixity: S&P members contributed significant input to the NDSA publication, *Checking Your Digital Content: What is Fixity and When Should I be Checking It?* [7]

Levels of Preservation: S&P members contributed significant input to the NDSA publication, *Levels of Preservation.* [8]

Conference Participation and Knowledge Sharing: S&P members made presentations about the *Levels of Preservation* at IS&T Archiving 2013, the 2013 NE NDSA Regional Workshop, the 2013 SAA Annual Conference, and iPRES 2013; and presented at Digital Preservation 2014 on *Checking Your Digital Content: What is Fixity and When Should I be Checking It?*

Wikipedia's Digital Preservation Entry: S&P members initiated a project to improve Wikipedia's coverage of digital preservation in general, but particularly in areas related to digital preservation terms, concepts, theories, strategies and history; standards, best practices and common methods; preservation repository architecture, operations and policies and certifying the trustworthiness of preservation repositories. While this Wikipedia article continues to evolve with recent contributions, S&P members helped frame the article by developing a new outline for the digital preservation article and improve the resources and citations.

In addition, the S&P periodically organizes "conference recap" sessions to expand the peer-to-peer network and share experiences and knowledge gained at conferences such as the Library of Congress Storage meeting, Research Data Alliance, iPRES, AMIA Hack Day, ICA, DLF Fall Forum and IASA. S&P members also periodically share updates on projects they are working on outside of the NDSA which have included the

UDFR, the Academic Preservation Trust, data management guides, DPN, repository self-assessments, a Drupal-based TRAC tool and many different institutional projects.

4. NEXT STEPS: FUTURE AGENDA

The S&P has identified several future projects including:

- Criteria for evaluating data repositories
- Metadata for complex objects to be emulated
- Updates on PREMIS 3, coverage of PBCore, EBU Core, SMPTE Core, and Bit Curator's DFXML.
- Standards adoption
- Preservation terms of service / SLAs
- Sustainability of our tools
- Preservation of social media

5. ACKNOWLEDGEMENTS

As a community driven organization, S&P is an organization of peers. The poster creators wish to thank all current and former S&P members and co-chairs, other NDSA Working Group members and invited speakers for their contributions to our collaborative efforts.

6. REFERENCES

[1] National Digital Stewardship Alliance (NDSA). 2015. Standards and Practices Working Group. Accessed: September 18, 2015. http://www.digitalpreservation.gov/ndsa/working_groups/standards.html

[2] National Digital Stewardship Alliance. 2015. Accessed: Accessed: September 18, 2015. http://www.digitalpreservation.gov/ndsa/index.html

[3] National Digital Information Infrastructure and Preservation Program (NDIIPP). 2015. Accessed: September 18, 2015. http://www.digitalpreservation.gov/

[4] Murray, K. June 29, 2015. We did all that? NDSA S&P project recaps. In *The Signal*. Accessed: September 18, 2015. http://blogs.loc.gov/digitalpreservation/2015/06/we-did-all-that-ndsa-sp-project-recaps

[5] NDSA Standards and Practices Working Group. February 2014. The benefits and risks of the PDF/A-3 file format for archival institutions. Accessed: September 18, 2015. http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_PDF_A3_report_final022014.pdf

[6] NDSA. September 2014. 2015 National agenda for digital stewardship. Accessed: September 18, 2015. http://www.digitalpreservation.gov/ndsa/documents/2015NationalAgenda.pdf

[7] NDSA Standards and Practices and Infrastructure Working Groups. 2014. Checking your digital content: what is fixity and when should I be checking it? Accessed: September 18, 2015. http://digitalpreservation.gov/ndsa/working_groups/documents/NDSA-Fixity-Guidance-Report-final100214.pdf

[8] Phillips, M., Bailey, J., Goethals, A. and Owens, T. 2013. Elaborating on the NDSA levels of preservation. In *Proceedings of IS&T Archiving Conference 2013 (ARCHIVING 2013)* (Washington, DC, USA, April 5, 2013). Accessed: September 18, 2015. http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Levels_Archiving_2013.pdf

The retroTECH Program at the Georgia Tech Library: Digital Preservation through Access

Sherri Brown, Wendy Hagenmaier, Lizzy Rolando, Jody Thompson, Alison Valk
Georgia Tech Library
Clough Commons
266 4th Street NW
Atlanta, GA 30332-0900
+1 (404) 894-4579
sherri.brown@library.gatech.edu; wendy.hagenmaier@library.gatech.edu; lizzy.rolando@library.gatech.edu; jody.thompson@library.gatech.edu; alison.valk@library.gatech.edu

ABSTRACT

This poster outlines the retroTECH program at the Georgia Tech Library, an innovative model of digital preservation in which hands-on access and campus community engagement are at the forefront.

General Terms

Institutional opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice; Training and education.

Keywords

Digital preservation; access; personal digital archiving; community engagement; teaching and research; vintage technology; digital archaeology; hardware and software preservation.

1. INTRODUCTION

For decades, archives have often emphasized preservation over access and waited passively for collections to come to them. With digital archives, however, obsolescence and the rapid pace of change have made it increasingly apparent that everyone must be an archivist of their own materials and that access itself can facilitate preservation. Archivists must act and engage our donors and users in the now. These ideas were the impetus for the retroTECH program at the Georgia Tech Library and are the focus of our iPres 2015 poster.

Built on collaboration with a strong multidisciplinary community of campus supporters and designed by a team of librarians and archivists with diverse expertise, retroTECH is a public-facing program in which the Library partners with the Georgia Tech campus community to design the future by hacking the past. With the emerging retroTECH Lab as a home base, students, faculty, staff, and alumni can undertake hands-on research, peer-to-peer personal archiving, curricular activities, and outreach around vintage technologies.

The inspiration for retroTECH grew out of user research interviews conducted with faculty in 2013 in preparation for a Library building redesign, and since then, has gained significant momentum. The idea takes what archives around the world are doing behind the scenes with digital forensics and born-digital workstations, combines it with a hackerspace ethos, and brings everything out for public-facing access, empowerment, and engagement. With the retroTECH program, we aim to reimagine digital archives by offering our technologically-savvy patrons a chance to use vintage, forensic, and emulation equipment typically restricted to library staff, museums, and specialized collectors. In addition to these models of institutional digital preservation practice, retroTECH has also drawn inspiration from the success of access-focused programs such as the Media Archaeology Lab in the Department of English at the University of Colorado Boulder[1] and the Computer and Video Game Archive at the University of Michigan Library.[2]

retroTECH has not only become a much-needed new library service; it represents a rich, unusual alignment of the Georgia Tech community's interest in the history and future of technology with Library faculty members' professional interests in digital data, archiving, visualization, and preservation. retroTECH aims to bring archiving to the people--and the people to the archive. Our conference poster outlines the spaces and services we are developing as part of the retroTECH program, including activities to gather requirements, pilot ideas, and design a lab that will open in our renewed Library building in 2018.

2. LAB IDEATION AND DESIGN

As we have envisioned it, the retroTECH Lab will not only serve as a hands-on historical reference point; it will activate new ideas about future technology and preserving innovation. The Library and Archives acquired our seed collection of five vintage workstations from the alum and former faculty member whose interview inspired the idea. Along with two emulation workstations, currently in development, these machines form the core of our pilot retroTECH Lab space, where we are testing programming to be implemented in our future permanent lab in the renewed building.

The vision for the retroTECH Lab entails a highly curated combination of classic, vintage hardware and software and cutting edge modern tools for emulation. Hands-on access, rather than preservation of the materials as museum objects, will be the main driver behind our collecting. We believe the benefits of easy

[1] http://mediaarchaeologylab.com/

[2] http://www.lib.umich.edu/computer-video-game-archive

access and open, experiential learning outweigh the potential risks of damage to the equipment. Our poster delineates the kinds of technologies--both vintage and new--that we envision in the program going forward. We also highlight the objectives of the program, including our belief that retroTECH will foster the kind of hacking that makes connections between the classic and the cutting edge and how we both engineer and are engineered by devices. The retroTECH Lab will share space with the Visualization Lab in our renewed library, further underscoring the links between past and future, between hardware and software, and between the material world and the virtual landscape. The poster presents graphical representations of our lab prototyping and ideation activities to date.

3. CURRICULAR PARTNERSHIPS, RESEARCH, AND PEER-TO-PEER ARCHIVING

Curricular partnerships are another crucial part of the retroTECH program. In spring 2015, the Library partnered with two instructors in the Writing & Communication Program to develop and implement a retrogaming assignment for six communication classes that were focusing on narrative in videogames. 132 students played computer games from the 1990s and early 2000s on the retroTECH computers and then prepared a blog post reflecting on the experience. Students also completed a post-gaming survey regarding their retroTECH experience and interests. Our poster highlights the feedback and ideas received from students who engaged in the retrogaming experience. We hope to expand our retroTECH curricular engagement moving forward to include working with classes in disciplinary areas such as the College of Computing, and we indicate other planned and potential opportunities for course-integrated use of the retroTECH technology.

We also are expecting to support faculty and researchers at Georgia Tech using the retroTECH equipment, and this section of the poster describes areas of research interest identified through user research. Possible areas of study might include historical hardware and software engineering, media archaeology, the evolution of game creation, emerging software development, and more. Furthermore, we aim for retroTECH to serve as a platform for users to conduct peer-to-peer digital archiving, working together and combining diverse expertise in order to recover, access, emulate, and preserve materials needed for research, for personal digital archiving, or for donating to the Georgia Tech Archives. We hope to create a cultural mindset that emphasizes the importance of archives, digital heritage, and long-term thinking, and to connect with the potential donors of born-digital collections that will fuel the Georgia Tech Archives' collection development strategy and attract researchers from across the country.

4. OUTREACH

In addition to academic and archiving partnerships that strengthen relationships with the teaching and research community on campus, connections with the vintage technology community have become just as important in our efforts to develop the retroTECH program. Community groups interested in exploring the history and evolution of technology are both potential partners in programming and sustainability and also sources of inspiration for our work moving forward.

In spring 2014, we presented at the Vintage Computer Festival Southeast alongside many vintage technology aficionados on the benefits of vintage computing in a library environment. Our poster will elaborate on our efforts to forge alliances with the regional historical computing groups and organizations behind this event and others, such as Maker Faire Atlanta, many of which are fueled by Georgia Tech alumni. Partnerships with these communities could facilitate volunteer opportunities, collection development and maintenance expertise, donations of equipment or archival collections, workshops, and support for the retroTECH program and the Library.

In fall 2014, we turned our outreach inward, curating a crowdsourced, rotating exhibit in the Library of vintage tech loaned by faculty, staff, students, and alumni. Fans of the exhibit cast over 3,100 online votes for their favorite items. Through the exhibit, we established a retroTECH Interest Group of over 100 people from all six academic units on campus, started a listserv dedicated to vintage technologies at Georgia Tech, and distributed a Community Engagement Survey to generate programming ideas. This section of the poster outlines the survey results to date, along with techniques for building community and sustaining momentum in the early stages of a new initiative, through innovative user-centered participatory programming.

We also detail outreach efforts planned for the pilot phase of the retroTECH Lab, including a partnership with the College of Computing on events for their 25th anniversary, and future outreach ideas, such as collaborating with an academic department to establish a permanent student assistantship for the Lab and partnering with other institutions on grants and events.

5. SIGNIFICANCE FOR THE DIGITAL PRESERVATION COMMUNITY

Through our poster outlining the retroTECH program, we aim to offer a model of digital preservation where access is at the forefront. We hope to illustrate how the traditional activities and expertise of digital preservationists can be enriched through participatory programming, a hands-on hacking mindset, and a peer-to-peer culture defined by long-term thinking. The poster offers the community a chance to learn from the challenges and successes of the retroTECH program to date, invites collaboration with digital curators working on allied projects, and serves as an inspiration to institutions hoping to establish similar programs.

6. REFERENCES

[1] Media Archaeology Lab, Department of English, University of Colorado Boulder, http://mediaarchaeologylab.com.

[2] Computer and Video Game Archive, University of Michigan Library, http://www.lib.umich.edu/computer-video-game-archive.

The Oracle Cloud Storage Archive for Long-term Storage and Preservation

Pyounguk Cho
Oracle
200 Oracle Parkway
Redwood City, CA 94065
650-506-5297
pyounguk.cho@oracle.com

Art Pasquinelli
Oracle
500 Oracle Parkway
Redwood City, CA 94065
650-607-0035
art.pasquinelli@oracle.com

ABSTRACT

Oracle Public Cloud offers data backup and archive services with a high level of data availability and at an affordable cost. This demonstration highlights key capabilities for direct usage of Oracle Storage and Archive Services via Openstack Swift API's and NFS interfaces. In addition, it also showcases integrated solutions using storage management tools such as iRODS and CommVault. A review of how both the technical and economic features of Archival Cloud Computing can be employed in new preservation infrastructure modeling, use cases, and academic and business scenarios will be given. The Oracle Archival Cloud offers new infrastructure opportunities to institutions.

1. INTRODUCTION

Oracle Storage Cloud Service is a secure, elastic, reliable, and cost-effective public cloud storage solution. It can be accessed from anywhere, 24/7, and from any device connected to the Internet. With zero investment in hardware, organizations can buy just as much enterprise-grade storage capacity as needed today, and buy more as required.

Oracle Storage Cloud Service provides an easy-to-use solution to store, manage, and consume large amounts of unstructured data over the Internet. Applications can access Oracle Storage Cloud Service programmatically by using either an OpenStack Swift-compatible REST API or Java API. Academic, library, and IT administrators can monitor key storage metrics and manage users and roles by using a web-based graphical console. Users can apply role-based access control for data stored on Oracle Storage Cloud Service at a very granular level. As required, data can be made accessible publicly.

Data that is stored using Oracle Storage Cloud Service is replicated on multiple storage nodes, guaranteeing protection against hardware failure and data corruption. Data is never moved out of the data center without owner permission. Oracle Storage Cloud Service can be employed as a cost-effective, remote backup solution for departmental, library, or enterprise data and applications. By backing up data and applications to Oracle Storage Cloud Service, users can avoid large capital and operating expenditures in acquiring and maintaining storage hardware. By automating backup routine to run at scheduled intervals, users can further reduce the operating cost of running a backup process. In the event of a disaster at your site, the data is safe in a remote location, and you can restore it quickly to your production systems. To learn more about Oracle Storage Service and to request for a trial subscription:

cloud.oracle.com/storage.

2. ARCHIVING AND PRESERVATION FEATURES

2.1 Security

Oracle uses enterprise-grade processes and operations to secure your data. For enhanced security, you can use the client-side encryption feature of the Java library. A cycle of encryption and decryption ensures that your data remains secure in the cloud. When customers use the client-side encryption feature of the Java library, for every object that is created in Oracle Storage Cloud Service, a unique symmetric key is generated. The Java library uses this key to encrypt data before storing it. After encrypting client data, the Java library encrypts the symmetric key as an envelope key by using an asymmetric key pair that the client provides. The envelope key is then stored as metadata alongside the object data. When customers use the Java library to access such encrypted objects, the envelope key is first retrieved and decrypted by using the asymmetric key pair that a user provides. The resulting symmetric key is then used to decrypt the object data.

2.2 Data Integrity

When an object is created in Oracle Storage Cloud Service from an uploaded file, the service returns the MD5 checksum of the object. This is in the ETag header of the HTTP response. The client that initiated the backup can verify whether the file was uploaded correctly by comparing the MD5 checksum provided by the service with a locally calculated checksum. Every request to Oracle Storage Cloud Service receives an HTTP response containing a status code, which indicates whether the requested operation was completed successfully. The client that initiated the backup can determine whether the data was backed up reliably, by interpreting the status code returned by Oracle Storage Cloud Service.

2.3 Authentication

Oracle Storage Cloud Service authenticates all requests through an authentication token mechanism. Every request to the service must include a valid authentication token, which the service provided previously in response to an authentication request containing a valid user name and password. The authentication token expires after 30 minutes.

2.4 Back Up Architecture

To facilitate the efficient and reliable upload of files that are larger than 5 GB, Oracle Storage Cloud Service supports uploading files in segments. This feature is called dynamic large objects. Users can segment a large file into multiple small files, each called a segment and each smaller than 5 GB, and then upload the segments individually to Oracle Storage Cloud Service. Customers must also create a manifest object, which will be used when the objects are downloaded, to concatenate the retrieved segments in the correct sequence and stream them in a single response. Note that customers can use their own convention-based schemes for segmenting large files.

Each operation on Oracle Storage Cloud Service is atomic. It either succeeds completely or fails completely. If the upload of a particular file fails, due to a network problem for example, the file must be uploaded again. Data that was uploaded until the network failure occurred is not saved in the cloud. So before you upload large files, even those that are smaller than 5 GB, consider segmenting them and then uploading the segments individually. With this approach, if the upload of a segment fails, only that segment needs to be uploaded again.

To optimize the storage space used in Oracle Storage Cloud Service, consider compressing data before uploading it. When this is done, data will consume less space in Oracle Storage Cloud Service and will take less time to upload and retrieve. A customer can store multiple directories and files in Oracle Storage Cloud Service with a single request, by packaging and compressing them and uploading the resulting tar.gz or tar.bz2 file.

In Oracle Storage Cloud Service, a container is created for each top-level directory and an object is created for each file.

2.5 New Economic Efficiencies

Oracle Archive Storage Cloud Service provides storage for applications and workloads that require long-term retention at the lowest price in the industry. As a “deep cloud” archive, the Archive Storage Cloud is suited for infrequently accessed large-scale data sets. It is priced at $0.001 GB/Month which equates to $12,000/PB/Year.

2.6 References and Citations

[1] Oracle eBook: Backing Up Data and Applications Securely, Reliably, and Efficiently (September 22, 2015)
Oracle Webpage: https://cloud.oracle.com/storage

Achieving Transparency and Replicability: A Data Curation, Verification, and Publication Workflow

Thu-Mai Christian
Odum Institute for Research in Social Science
University of North Carolina at Chapel Hill
thumai@email.unc.edu

Sophia Lafferty-Hess
Odum Institute for Research in Social Science
University of North Carolina at Chapel Hill
slaffer@email.unc.edu

ABSTRACT

In this poster, we illustrate the workflow developed by the Odum Institute for Research in Social Science Data Archive to support the curation and verification of replication data files for the *American Journal of Political Science*.

General Terms

Preservation strategies and workflows

Keywords

Data curation, Data quality, Replication, Verification

1. POSTER SUMMARY

In a move to promote open scientific inquiry, several major journals have issued policies requiring authors to make the data underlying results presented in their published articles openly available to others, which enables verification and replication of findings. In doing so, the journals protect the integrity of the scientific record [4] while also enhancing the visibility and impact of research [5]. Unfortunately, submission of dataset files to a repository has not been adequate to ensure the long-term preservation and reuse of these data for these purposes.

In the frequently cited article, "Replication, Replication," Gary King asserts that the "replication standard holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party could replicate the results without any additional information from the author" [3]. In an effort to uphold King's replication standard, the editorial staff of the *American Journal of Political Science (AJPS)* recently issued a revision to its data availability policy, which had already required authors to upload replication files to a designated open access repository prior to submission of the final manuscript for publication [1]. Despite this initiative, the *AJPS* editorial staff has recognized the varying quality of replication files currently housed in the *AJPS* data repository [2]. In response to this, the new *AJPS* replication policy revision stipulates that article publication is contingent not only on the submission of supporting files—including the data, programming code, codebooks, and other explanatory text—but also the successful replication of tables and figures in the final manuscript using the submitted files [1].

The Odum Institute for Research in Social Science Data Archive has been tasked to perform the verification of replication datasets and ensure the comprehensiveness of submissions. Even with guidance provided to authors on how to prepare replication files, the quality of data submissions has varied, with only a fraction able to reproduce tables and figures in final manuscript drafts on the first attempt. Missing codebooks, incomplete or non-commented programming code, rounding errors, mismatched figures, and an array of other issues have added complexity to both the publication and data curation and verification workflows. Because of this, it has been necessary to develop a standard, integrated workflow that relies significantly on cooperation of and coordination between the author, editor, and data archive in order to ensure that submitted files meet quality standards for both replication and preservation and reuse.

This poster will outline the human-driven workflow to archive, verify, and link replication data to their associated journal publications, as well as its integration into the scholarly publication workflow. The poster will also describe critical issues, key lessons, and potential opportunities for archives working to preserve scholarly assets to help sustain the research enterprise.

2. ACKNOWLEDGMENTS

The authors would like to thank the editorial staff of *AJPS* and the editor of *AJPS*, Dr. William Jacoby, for their assistance in the development and implementation of the data curation, verification and publication workflow. We would also like to thank the statistical consultants at the Odum Institute who perform the meticulous task of verifying that the submitted replication files can successfully reproduce the findings in final manuscripts. They play a key role in achieving the goal of research transparency and replicability.

3. REFERENCES

[1] Jacoby, W. (2015, May 26). The AJPS replication policy: Innovations and revisions. Retrieved from http://ajps.org/2015/03/26/the-ajps-replication-policy-innovations-and-revisions/

[2] Janz, N. (2015, May 4). Leading journal verifies articles before publication--So far, all replications failed [Blog]. Retrieved from https://politicalsciencereplication.wordpress.com/2015/05/04/leading-journal-verifies-articles-before-publication-so-far-all-replications-failed/

[3] King, G. (1995). Replication, replication. PS: Political Science & Politics, 28(3), 444. http://doi.org/10.2307/420301

[4] National Research Council & Committee on National Statistics. (1985). Sharing research data. (S. E. Fienberg, M. E. Martin, & M. L. Straf, Eds.). Washington, D.C.: National Academy Press. Retrieved from http://www.nap.edu/openbook.php?record_id=2033

[5] Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. http://doi.org/10.7717/peerj.175

In Search of GeoBlacklight: Reporting on a Community-Driven Geospatial Data Portal in the Library

R. Shane Coleman
University Libraries, Virginia Tech
560 Drillfield Drive
Blacksburg, VA 24061
+1 540-231-8665
shanecoleman@vt.edu

Andrea L. Ogier
University Libraries, Virginia Tech
560 Drillfield Drive
Blacksburg, VA 24061
+1 540-231-9255
alop@vt.edu

Mohamed M.G. Farag
Computer Science Department
Virginia Tech
Blacksburg, VA 24060
+1 540-449-9019
mmagdy@vt.edu

ABSTRACT

Geospatial data are widely used by many institutions, governments, and corporations; given the diversity of organizations concerned with geospatial data, preserving and curating these important digital files presents unique challenges to the preservation community. In the University Libraries at Virginia Tech a small project team is working to use a local implementation of GeoBlacklight to build a flexible geospatial data portal that addresses the needs of diverse stakeholders on campus, including those outside the context of academic research. This poster will present the results of an initial geospatial data assessment, the issues and concerns of each stakeholder on campus, and how the GeoBlacklight implementation addresses both the challenges posed by the stakeholders and by the complexity of geospatial data itself.

The Data Curation Unit in Newman Library at Virginia Tech is creating opportunities for geospatial discovery and preservation through collaboration with University Facilities and University IT. Libraries have long been known for institutional repositories that lack storage capacity and have outdated features. Through collaborative efforts with University Facilities and University IT the library has started implementation of an instance of GeoBlacklight to resolve these issues for our geospatial data users. The Facilities group has provided a variety of dataset use cases along with metadata schema input. IT has been responsible for setting up server space which allows for deposits of more than two gigabytes. The library has been responsible for the development of the interface. With this team we aim to provide a secure environment that incorporates the needs of non-academic patrons alongside a more traditional data repository, making it easier for campus users to deposit and extract data, and store larger sets of data than has previously been possible. This poster will present the results of our initial geospatial data assessment, the obstacles posed by working with numerous university stakeholders, and what we believe to be a sound solution for geospatial data discovery and preservation of both academic and non-academic geospatial data at Virginia Tech.

Geospatial data generated for practical use (not specifically for research) became of interest once we began partnering with our University Facilities and University IT groups. We found that these campus departments frequently produce geospatial data that is often of interest to research groups across campus. In addition, there are several groups on campus that collect historical state government data containing geospatial components. Beginning with the acknowledgement that all data should be discoverable, we developed workflows that enable our GeoBlacklight instance to treat research and non-research data the same. Essentially we are creating a single discovery platform for all geospatial information acquired and created by Virginia Tech.

We also found significant benefit in the collaboration with University IT, including access to greater storage space and stricter backup protocols. This partnership was developed on the premise that the library would help curate and make discoverable all geospatial data in use at Virginia Tech, as long as they University IT was responsible for maintaining server space and managing the backup processes for the datasets in their care. This allowed us to have experts focus on the storage and management of geospatial data allowing library staff to focus on collection and curation of the data.

In essence, the GeoBlacklight project has become a means for better aligning the library as the central location for finding and accessing geospatial data at Virginia Tech, and has also allowed us to leverage campus partners to better optimize our ability to serve the discovery and preservation needs of the Virginia Tech community.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges;

Keywords

Geospatial data, GIS, GeoBlacklight, Hydra.

1. REFERENCES

[1] Hardy, D., and Durante, K. 2014. A Metadata Schema for Geospatial Resource Discovery Use Systems. *Code{4}Lib Journal* 25 (July 2014). http://journal.code4lib.org/articles/9710

[2] McGarva, G., Morris, S., and Janee, G. 2009. Preserving Geospatial Data. *Technology Watch Report 09-01.* Digital Preservation Coalition. http://www.dpconline.org/component/docman/doc_download/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-gred-greg-janee

Preserving In-House Developed Software

Nicole Contaxis
National Digital Stewardship Resident
National Library of Medicine
Bethesda, MD
301-443-8000
nicole.contaxis@nih.gov

ABSTRACT

Software often plays a key role in the ways that institutions function, and some institutions, like the National Library of Medicine (NLM), have a history of developing software to serve their unique needs. Preserving this in-house developed software is the goal of the National Digital Stewardship Residency Project, "NLM-Developed Software as Cultural Heritage." This project will not only help ensure access to content created on this software, but, in the case of NLM, will also help document a long and often unrecognized intellectual history [1]. Although copyright concerns are largely avoided since the software was produced in-house, a variety of administrative obstacles still need to be addressed before the technical process of software preservation can begin. These obstacles include but are not limited to: (1) locating knowledge sources for software projects that are long defunct; (2) locating usable copies of software, either tangible or intangible, that may not have been properly documented or stored; and (3) tracing the history of projects that may have gone through several re-branding efforts or versioning's. This poster will address these issues as they have affected the current project at NLM and will demonstrate how a properly conducted inventory is necessary for contending with these obstacles and ensuring a reliable long-term software preservation strategy.

1. BACKGROUND

The National Library of Medicine has been developing software for internal and external use since the early 1960's when they began work on the Medical Literature Analysis and Retrieval System (MEDLARS). This computerized bibliographic system was meant to facilitate access to the library's bibliographic and serial records and to help compile the extensive indexes being produced at the time. With the creation of GRACE (Graphic Arts Composing Equipment), a custom phototype-setting machine, NLM was able to provide access to their data and print their indexes in record time. GRACE is now housed at the Smithsonian Institution [2].

In the 1970's with the creation and implementation of MEDLINE, NLM made their data available online, and years later, began to work on ways to offer full-text access through a variety of networks. In conjunction to their bibliographic systems, NLM experimented with a range of ways to satisfy the information needs of the health services community at large. Examples of such experiments include satellite communication to assist physicians in remote areas, the first internal library system, mobile computerized workstations to assist workers at toxic waste spills, user-friendly interface software for bibliographic access, and search and indexing features for GenBank and other influential databases [3].

The National Digital Stewardship Residency project aims to help preserve this history and call attention to the importance and impact of software development both within the library and beyond. The poster will represent a key part of this project as it illustrates how to deal with administrative issues using a thorough inventory before beginning the technical process of preservation.

2. OBJECTIVES

As this brief history of software development at NLM illustrates, software functions can vary wildly. Some projects will have user-interfaces while other will only perform computations. In order to make informed choices about how to preserve a software project, it is necessary to learn about that piece of software, how users interact or interacted with it, how it was developed, and how it changed over time. With the range of software projects from the history of NLM in mind, a delicate approach to the contexts of each project is necessary before committing to a preservation strategy for a particular piece of software. If, for example, a piece of software is meant to help users access and interact with information, it may be best to create an emulation in order to preserve it adequately. However, if a piece of software is important because of its computational uses, migration may be a better option. Furthermore, these types of decisions may be affected by the way a piece of software changes over time. Change occurs frequently for software projects, as versions and patches can affect the overall nature and experience of a project.

This poster intends to outline what information is necessary to inform these sorts of decisions, provide an example of how an inventory can be compiled, and illustrate how the peculiarities of software development can be accommodated in an archival setting. The inventory process necessitates close attention to the institution's history as well as to the ramifications of different preservation tools and techniques on the longevity of and meaning associated with a software project.

3. ACKNOWLEDGEMENTS

Funding and support for this National Digital Stewardship Residency comes from the Library of Congress and the Institute of Museum and Library Services

4. REFERNCES

[1] Library of Congress. 2015. National Digital Stewardship Residency. Webpage. http://www.digitalpreservation.gov/ndsr

[2] C.R. Dee, "The Development of the Medical Literature Analysis and Retrieval System (MEDLARS)," *Journal of the Medical Library Association.* 2007 Oct 95(4): 416-425.

[3] B. L. Humphreys, "Adjusting to progress: interactions between the National Library of Medicine and health science librarians, 1961-2001," *Journal of the Medical Library Association*. 2002 Jan 90(1): 4-20

Addressing Major Digital Archiving Challenges

Dr Janet Delve
University of Portsmouth
Eldon Building, Winston Churchill Avenue
Portsmouth, PO1 2DJ, UK
Janet.Delve@port.ac.uk

Professor David Anderson
University of Portsmouth
Eldon Building, Winston Churchill Avenue
Portsmouth, PO1 2DJ, UK
David.Anderson@port.ac.uk

Dr Andrew Wilson
University of Portsmouth
Eldon Building, Winston Churchill Avenue
Portsmouth, PO1 2DJ, UK
Andrew.Wilson@port.ac.uk

ABSTRACT

The E-ARK project (E-ARK is funded by the European Commission's FP7 PSP) is addressing several major challenges faced by archives and institutions/researchers preparing data to send to archives. With the recent emphasis on open access, there has been a sea-change regarding discovery and archival material, so that citizens, businesses and academic researchers as well as the archives and data providers themselves can look forward to novel ways of analyzing archival data. E-ARK is half way through its three-year timespan, and has already produced some concrete solutions to real challenges in this problem space. This poster will graphically demonstrate the various challenges and show how E-ARK is meeting them now, or plans to in the future.

General Terms

Infrastructure opportunities and challenges; Frameworks for digital preservation.

Keywords

Digital Archives, User Survey, E-ARK, EC, ICT-PSP, Pilot, e-infrastructure, data mining, OAIS, Big Data, born-digital records, ingest, access, EDMRS, database preservation, open access, MoReq.

1. INTRODUCTION

By the time of iPres 2015, the E-ARK project will have gone past the halfway point, and have started producing key results and drafts that are relevant to the digital archiving problem space. It is timely to share these results with the wider digital archiving community.

Our focus in essence: in the first year we carried out a suite of best practice surveys to see how various communities carry out basic digital archival tasks: preparing their archival material, storing it, and subsequently accessing it. We also set up a knowledge centre proof of concept to house the expertise gathered over the life of the project. In the second year, we are developing draft standards and open source tools to perform these tasks, underpinned by a comprehensive legal study, which provides a European legislative backdrop for the work we are doing. In the third year, we will have an integrated framework with modular components that will be deployed in seven pilot instances. These will cover a range of data types, archival institutions and discovery methods. Included in this will also be a data mining showcase based on Big Data methods that have been used to help develop the framework. The pilots will be based on real use cases that can serve as exemplars for many other user communities.

2. ABOUT E-ARK

European Archival Records and Knowledge preservation (E-ARK) was launched in February 2014 and is a 3-year pilot project within the European Commission's ICT Policy Support Programme (PSP) Competitive and Innovation Framework (CIP) Pilot B Programme under Grant Agreement no. 620998. With 16 partners in 11 EC countries comprising end users, research institutions and systems suppliers, its objective is to provide a single, scalable, robust approach capable of meeting the needs of diverse organisations, public and private, large and small, and able to support complex data types. E-ARK will demonstrate the potential benefits for public administrations, public agencies, public services, citizens and business by providing simple, efficient access to the workflows for the three main activities of an archive - acquiring, preserving and enabling re-use of information.

E-ARK will implement a number of pilot systems in different countries addressing challenges which differ in content and scale in order to create, by the end of the project, in 2017, a suite of openly-accessible end-to-end solutions capable of integration into third-party products and which will be sustained into the future.

Our work is worldwide: the first attempt to bring together working elements of archival systems. As such it is an ambitious project which has several key features: creating standardized pre-ingest formats / specifications; expanding MoReq modules to be used as a key element of the infrastructure; using CMIS and Big Data techniques to promote new ways of access to digital archives, etc. It also addresses a wide range of users: public bodies, commercial institutions, individual citizens and researchers.

Our project will also provide a Digital Preservation Maturity Model which will enable organizations to not only assess their current performance, but also to measure improvement. More information about the project is available from our website at www.eark-project.eu.

3. WHAT ARE THE KEY CHALLENGES?

Here is a list of major challenges in this domain, from a variety of perspectives and communities, with descriptions of the E-ARK approach to addressing these challenges:

- *How do I get my data out of my electronic records system (e.g. Sharepoint) and into an archive?*
  This is a major priority for E-ARK, as it can be a real headache for e.g. government departments to package their data for transfer into an archive in the manner that the archives require. Practical issues include the fact that records systems use their own hierarchical classifications, which do

not necessarily match those used by an archive. How can such compatibility problems be overcome? We have studied current best practice, and are now working on data export specifications that data producers can use to get their data out of their source systems and into the archives in an "archive-acceptable" format. These open specifications are based on the MoReq schema, and cater for a wide range of systems, including Electronic Records Management Systems (ERMSs) as well as simple file based systems. Our Work Package 3 (WP3), led by the National Archives of Estonia, is spearheading this work, and they have already produced reports, conference presentations etc. with key practical information.

- *How do I archive databases?*
  E-ARK is producing everything you need for each stage of digital archiving, and we are covering database archiving as well as the archiving of digital records. We have studied current best practice, and based on this we have produced draft specifications showing how to put data (including databases and their contents) into an archive, store them there, and then access them later for discovery and re-use. We have been working closely with the Swiss Federal Archives and the Swiss Koordinationsstelle für die dauerhafte Archivierung elektronischer Unterlagen (KOST), and you will find the latest version 2.0 of the Software Independent Archival of Relational Databases (SIARD) format on our website for your feedback. Our Work Package 4 (WP4), led by the Austrian Institute of Technology, is spearheading this work, and they have already produced reports, conference presentations etc. with key practical information.

- *Are there any general models or schemas that show the digital archival processes step by step?*
  We have produced a comprehensive general model that is fundamental to our entire project: it covers all the tools, processes, workflows, users etc. and specifically includes the pilot implementations (various parts of our final E-ARK system will be piloted by 7 national archives). Our Work Package 2 (WP2), led by the National Archives of Hungary, is spearheading this work, and they have already produced reports, conference presentations etc. with key practical information

- *Are there any new ways to discover archival data? Can we do complex searches or just google type searches?*
  We are looking at new ways of discovery for a wide range of data and many types of users: businesses, researchers, citizens, government departments etc. Whilst sensitive data has to be protected, we are looking for the best tools and techniques for accessing and analyzing any data that is open for discovery. We have studied current best practice in this area, and have used our findings as a basis for our developments which include data mining, Online Analytical processing (OLAP) and other advanced searching techniques. Our Work Package 5 (WP5), led by the Danish National Archives, is spearheading this work, and they have already produced reports, conference presentations etc. with key practical information. Our Work Package 6, led by the Austrian Institute of Technology, is also contributing to the advanced searches effort with a report on faceted searches.

- *What has Big data got to do with digital archiving? What is Hadoop and can we use it?*
  Big Data is a broad church, but can be said to involve powerful (fast) analysis of large volumes of varied data to produce valuable new insights with greater accuracy. Big Data has associations with open data, and cloud computing, with an emphasis on large-scale accessibility. The Apache™ Hadoop® project develops open-source software for reliable, scalable, distributed computing. We have developed an integrated system using a Hadoop cluster, running software such as Solr, Hive, Pig, Mamout etc. Big Data also leans heavily on previous architectures such as multi-dimensional databases and data warehousing, and we are using Big Data techniques such as data mining, data warehousing, dimensional modelling and Online Analytical Processing (OLAP) to carry out large-scale analysis of e.g. geographical data (geo-data) using Oracle Warehouse Builder and Oracle OLAP. Our Work Package 6 (WP6), led by the Austrian Institute of Technology, is spearheading the Big Data work, and they have already produced reports, conference presentations etc. with key practical information. Work Package 5 (WP5), led by the Danish National Archives, is using Big Data techniques for discovery, and they have already produced reports, conference presentations etc. with key practical information. Work Package 4 (WP4), led by the Austrian Institute of Technology, is using Big Data techniques such as dimensional modelling to archive databases, and they have already produced reports, conference presentations etc. with key practical information.

- *Are there any standards to help me archive my data properly?*
  Producing specifications and schemas forms a vital part of our work, alongside developing open source tools / workflows and frameworks. Standardizing the digital archival process across Europe and beyond should be a real help to institutions large and small, governmental, commercial or academic. We are creating our schemas to be as flexible and useful as possible – with mandatory elements that are essential to comply with best practice, and plenty of flexible options so that institutions / individuals can customize their archives in myriad different ways. Our work is not just for national archives – we do everything with regional and local archives in mind too. We have several reports dealing with standardizing issues

- *How does digital archiving vary from country to country in Europe?*
  We have a broad range of national archives taking part in E-ARK, with many more countries represented in our Archival Advisory Board. This enables us to take account of many different types of archival practice: some archives currently have no digital archives, some archives deal with everything as a database, some archives deal only with records etc. We covered current archival practice across Europe in our best practice reports in the first year of the project.

- *How does the law affect digital archiving in each European country? Are there any EC laws / directives that affect all digital archiving, and what is on the horizon in this respect?*
  These are vital considerations for E-ARK as each country needs to be able to use our outputs within their own legal framework. For this reason, we have undertaken comprehensive research to determine upcoming legislation that will affect practical digital archiving. We have a dedicated, extensive legal study which we will keep updated throughout the project.

- *Do the archives have any examples or use cases to inspire me?*

We are developing pilot cards to show how our archival partners will actually be using E-ARK in their pilot implementations. These cards will highlight the use cases for each national archive, showing why they joined the project and what benefits they expected to gain.

- *Are there any Open Source digital tools I can use? Can they be integrated? Will they fit with commercial tools / systems and existing Open Source tools / systems?*
  Our tools and platforms are all designed to be scalable and open source, so they will be suitable for your archiving needs. We have leading open source and proprietary commercial partners both in the project consortium, and on our Commercial / Technical Advisory Board, in order to ensure integration and a good fit with existing archival systems. Our aims with respect to scalability are covered in Work Package 6 (WP6).
- *Can I use something developed for a national archive in my regional archive/ local archive/ research data center?*
  Yes, this is our plan: our designs are for all archival shapes and sizes. We have representation from national and regional archives, and would also welcome input from local archives / research data centres.
- *How can I measure how well my organization is performing in terms of digital archiving? Are we beginners, or a bit further along the road?*
  We have been working on a specialized business maturity model to enable institutions to gauge their progress in this regard. All the information necessary for digital archiving, including vocabulary management, is going into a dedicated, long-term Knowledge Base, to be hosted by the DLM Forum on their website. Our Work Package 7 (WP7), led by the Instituto Superior Técnico, Lisbon, Portugal, is spearheading this work, and they have already produced reports, conference presentations etc. with key practical information.
- *I am unsure about or don't like something that E-ARK is doing. How can I make my thoughts known to them?*
  Please send us your feedback – email info@eark-project.eu!
- *Does it matter which preservation strategy I use with digital archiving? For example, can I use migration or emulation or a hybrid?*
  E-ARK is preservation-strategy neutral, and we are consciously identifying metadata (data about data) elements catering for both migration and emulation.
- *Do you have any questions for us? If so please get in touch. You can join our mailing list (http://eepurl.com/M35bH), and we are looking for more members on our Data Provider Advisory Board; and local archive members for our Archival Advisory Board (contact Andrew.Wilson@port.ac.uk).*

4. ABOUT THE AUTHORS

Dr Janet Delve, E-ARK Co-ordinator, University of Portsmouth.

Dr Andrew Wilson, E-ARK Senior Research Fellow and Advisory Board Co-ordinator, University of Portsmouth.

David Anderson, Professor of Digital Humanities, University of Portsmouth.

Establishing Trustworthy Repositories of Scientific Data: Opportunities and Benefits

Robert R. Downs
Center for International Earth Science
Information Network (CIESIN),
Columbia University
61 Route 9W
Palisades, NY 10964
+1 (845) 365-8985
rdowns@ciesin.columbia.edu

Ruth Duerr
Ronin Institute for Independent
Scholarship
6637 W 95th Pl
Westminster, CO 80021
+1 (303) 946-4842
ruth.duerr@ronininstitute.org

Devan Ray Donaldson
Department of Information & Library
Science
Indiana University
1320 E. 10th St, Wells Library 019
Bloomington, IN 47405
+1 (812) 855-9723
drdonald@indiana.edu

Sarah Ramdeen
School of Information and Library
Science
University of North Carolina at Chapel
Hill
Chapel Hill, NC, 27599
ramdeen@email.unc.edu

ABSTRACT

Scientific progress often depends on the ability of the scientific community to build on the works of others. Such works include scientific data, published reports of findings, and other research-related information and artifacts that are produced as part of the scientific process. Providing capabilities for accessing and using such scientific works enables the reproducibility of published methods and results to identify opportunities for improvement. Access and use of science products also enables others to build on previous work. In an increasingly digital world, the science community accesses and uses relevant scientific resources that have been obtained from digital repositories, data centers, and archives, as well as from traditional sources such as publishers of journal articles. Digital repositories need to establish capabilities, which provide access to and enable the use of digital resources. These resources are needed by the science community to improve and build on the efforts of others. Digital repositories that provide free and unrestricted access to scientific data and research-related information can reduce the barriers to science. By reducing these barriers they will be able to offer opportunities for members of the scientific community to pursue research questions and challenges that were previously unapproachable. These may include opportunities for researchers to gather data from other domains and support interdisciplinary research. Opportunities to use the data products and services offered by digital repositories also can contribute to the development of the scientific community and to the emergence of new areas of study.

Being able to access scientific data and other research resources supports future science and is important to the research community. Given their role as stewards, digital repositories must be considered by this community to be trustworthy. With limited resources available in many science domains, the scientific community as a whole cannot afford to lose science data and related resources. Digital resources may be particularly vulnerable to loss. Improving the infrastructure and practices for managing scientific data can reduce the potential for such losses. Trustworthy facilities are needed to curate, disseminate, and maintain these data and research-related materials. Furthermore, trustworthy repositories are needed to develop and improve data management services. They should also foster improvements in the capabilities and practices for scientific data stewardship.

Establishing digital repositories as trustworthy stewards of scientific data and related research products and services offers potential opportunities and benefits for science and society that can be leveraged to further research, educational, or decision making objectives. The direct beneficiaries of science data repositories include the individuals who serve as producers, stewards, and users of science data as well as the organizations that fund and host the digital repositories. Other potential beneficiaries include those who are not community stakeholders, currently, but may have an interest in these resources in the future. Furthermore, as described below, society at large also could benefit from digital repositories that have been recognized as trustworthy stewards of scientific data.

Data producers include scientists and other members of science project teams. Such data producers can compare digital repositories to determine those that are trustworthy, thereby enabling consumption of their data by interested researchers. If there is a variety of trustworthy repositories for their data, data producers can be selective and choose the repository that will best serve the community of potential users that the data producers are targeting for the use of their data products and services. Furthermore, selectivity among data producers for their choice repository could lead to competition among repositories that serve a particular science discipline, which may in turn lead to increased specialization of repositories to provide unique services.

Science data stewards include professionals in data management, information systems, and data services. These stewards can

compare choices among employment opportunities where they will apply their knowledge and skills, while also contributing to the curation, preservation, and dissemination of scientific data products and services. Tools should be available to enable data stewards to prepare, process, and preserve data for the future. These tools should also enable the dissemination of data products and services to diverse communities of users. Data stewards who accept positions at trustworthy science data repositories can enjoy the opportunities for professional development. These opportunities may be more abundant for individuals working within organizations that have been designated as trustworthy providers of scientific resources. Trustworthy digital repositories of science data and their staff should be able to demonstrate sustainable capabilities for managing data curation operations, for diligently preserving and disseminating science data, and for ensuring the integrity of their systems.

Users of science data products and services include researchers of all types, decision-makers, learners, and members of the general public. With improvements in the quality of products and services available from trustworthy digital repositories, these users can patronize trustworthy digital repositories that offer resources relevant to their goals and interests. Trustworthy repositories may recognize the potential for expanding their user base by supporting various levels of expertise, particularly in the science domain represented by the data that they archive and disseminate. As such, the availability of data products and services curated by trustworthy repositories for current and future users will enable use by increasingly diverse populations.

Benefactors of trustworthy digital repositories of science data will be able to observe and demonstrate how their support of trustworthy resources that have been prepared and disseminated, contributes to the future of science and the overall benefit of humanity. Funders that support the development and operation of trustworthy digital repositories could include government agencies, foundations, and philanthropists. The costs incurred enable the stewardship and dissemination of science data products and services. Perhaps, with such evidence of the value of their contributions, funders will be able to provide trustworthy digital repositories with the support needed to sustain science data stewardship operations and to help prepare science data repositories with the capabilities necessary to meet future challenges for the curation and dissemination of science data.

Organizations that host trustworthy digital repositories often include domain-specific data centers, archives, and institutional repositories. Upon attaining the status of trustworthy digital repositories, these organizations are likely to recruit the most qualified members of the science community based on their reputation for providing reliable data products and services. As more organizations become trustworthy digital repositories of science data, we also can expect the requirements of being trustworthy to increase. Just as other standards improve as technology evolves and new needs are recognized, the demands for increasing the requirements for trustworthy digital repositories, especially those that are responsible for science data, also should become more rigorous. This will ensure that science data and other research materials in digital form are being managed effectively for future use.

Members of society who are not traditional users of science data or other scientific resources also can benefit from the emergence of trustworthy science data repositories. Open science data that are accessible from trustworthy digital repositories will offer societal benefits as the data are used and translated into knowledge that contributes to the well-being of society at large. For example, educational institutions will be able to leverage the data and other research materials available to improve opportunities for educators and their students to learn from such resources. In addition, the benefits of trustworthy digital repositories for science can be realized by society as scientific breakthroughs, made possible by the continuing availability of science data products and services, thereby contributing to the lives of current and future generations as data are used to inform decision-making.

These are just a few of the opportunities and benefits that we can expect and hope for as digital repositories for science data attain the designation of being trustworthy. Taken together, the opportunities and benefits that can emerge from the availability of trustworthy digital repositories for science data can increase the maturity of the infrastructure and capabilities for managing, curating, disseminating, and preserving the digital data that exist today as well as those that will be produced in the future. Likewise, the availability of trustworthy science data repositories also has the potential to increase the professionalism of scientific data management practices, reducing the potential for the science data that have been created in digital form to be lost, through technological obsolescence, mismanagement, insufficient context for use, lapses in security, or other potential difficulties that could occur. Progress in the infrastructure for science can be achieved through the development and operation of trustworthy digital repositories for science data.

In summary, the availability of digital repositories that have been designated as trustworthy stewards for scientific data can contribute to the future availability of the science data that have been created during recent decades as well as the data that will be created in the future. Trustworthy digital repositories for science data can raise the expectations of stakeholder communities, increase the availability of choices for managing science data, improve scientific data stewardship practices, and contribute to the progress of science and the betterment of humanity.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges.

Keywords

trustworthy digital repositories; scientific data centers; science data infrastructure; cyberinfrastructure; data archives; science data products; data services.

1. ACKNOWLEDGMENTS

Support for Robert R. Downs has been received from the National Aeronautics and Space Administration (NASA) under Contract NNG13HQ04C for the continued operation of the Socioeconomic Data and Applications Center (SEDAC).

Alternatives for Long-Term Storage Of Digital Information.

Chris L. Erickson
Brigham Young University
HBLL 2217, Provo, UT 84602
1-801-422-1851
chris_erickson@byu.edu

Barry M. Lunt
Brigham Young University
265 CTB, Provo, UT 84602
1-801-422-2264
luntb@byu.edu

ABSTRACT

The most fundamental component of digital preservation is managing the digital objects in archival repositories. Preservation Repositories must archive digital objects and associated metadata on an affordable and reliable type of digital storage. There are many storage options available; each institution should evaluate the available storage options in order to determine which options are best for their particular needs.

This poster examines three criteria in order to help preservationists determine the best storage option for their institution: Cost, Longevity, and the Migration Time frame. While Richard Wright maintains that “storage is becoming the lowest cost in a digital repository,” Cost is probably the single most important factor when considering long term storage. Cost may be a limiting factor in the number objects that are preserved. Chapman asserts that repository storage costs “must be affordable and manageable or content owners will withhold materials from deposit.” Storage costs, even if they are declining, may influence decision makers to select a low-cost storage option, at the expense of essential preservation needs. DeRidder, in her presentation “Considerations for Storage and Protection of Content”, lists Cost as the first factor in choosing a storage media option.

Figure 1, included at the end of page 2, shows the costs of institutional storage, cloud storage, and alternative types of digital storage that we looked at when considering storage possibilities. (The author gathered this information from internet sources or from the storage providers directly).

Another very important criterion regarding digital preservation is the average lifespan of digital media. Selecting long-lived media for archiving digital content affects not only the end costs, but the long-term safety of the objects as well. Typical digital storage media have an expected lifespan of 3 – 10 years, though failure could occur at any time. Short-lived media, when combined with ineffective backup procedures, can result in the permanent loss of digital content.

Figure 2, Average Lifespan of Digital Media – Years, shows the often-quoted potential lifespan in years for different types of media. This figure also shows the realistic lifespan of the same media from our experience. Evidence compiled so far through observation and institutional experience show that the actual lifespan is often far below the advertised lifespan.

The only exception to this discrepancy is the M-Disc. This disc, developed at Brigham Young University and available in DVD and Blu-ray formats, is the only method that makes an irreversible physical change to a digital medium. The M-Disc is highly resistant to any of the normal factors that degrade digital objects, such as light, heat, humidity, temperature change, magnetism, bit rot and bit flips.

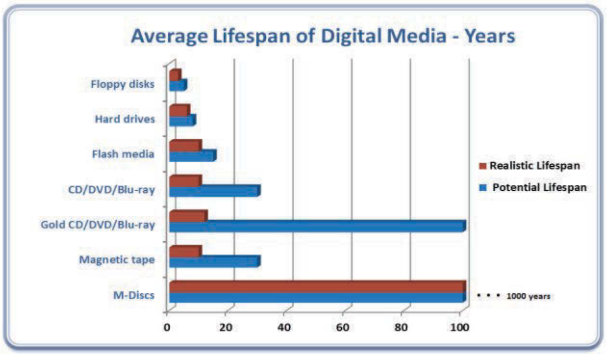


Figure 2: Average Lifespan of Digital Media – Years

The third criterion that involves both the lifespan of the media and the resulting costs of digital preservation is the Migration Time Frame. Every digital medium and every digital system has a limited lifespan. The media will eventually fail or the system will become obsolete. In order to preserve digital objects beyond the expected lifespan of the media, most digital media need to be refreshed or migrated regularly.

Figure 3 shows the expected migration time frame for hard drives, computer tapes, and the long-lived M-Disc optical discs. Each round of migration has an additional risk of data loss, and repeated migrations increase the probability of loss.

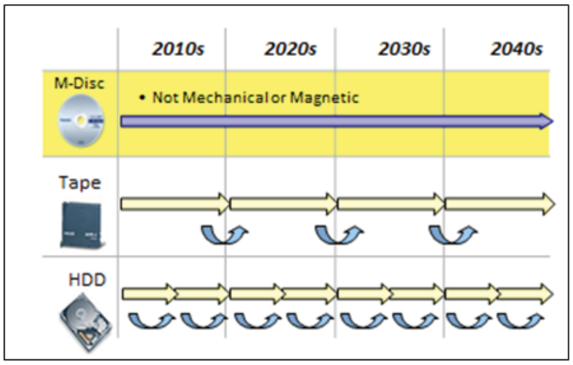


Figure 3: Migration Time Frame

The cost of migrating from one generation of media to another, or from one type of media to another type can be significant. Tape devices, such as with LT0 tapes, can only write to the current and one previous tape generation, and read the two previous generations. Thus, in three generations of LTO tapes, which currently is approximately ten years, the tapes and drives could become obsolete. Migration would include the costs of the new media and the new systems, but it must also include the personnel costs to manage the replacement and verification processes so that there is no data loss or degradation. Hard drives have a limited lifespan; they must be replaced regularly.

The Library of Congress archiving website recommends creating new media copies every five years or when necessary to avoid data loss. Since it is not possible to predict accurately when a drive or media will fail, it is important to refresh or migrate your digital media every few years. However, long-lived optical discs, such as the M-Disc DVD or Blu-ray, do not require refreshing or migration, thus adding to the cost savings.

Until recently, our institution stored full resolution digital collections in three ways: on optical discs (gold CDs and DVDs); on a variety of tape formats; and on external hard drives. Since each of these types of media experienced failures, multiple copies of every archived collection were required. Here are some problems we encountered with these media:

- Name brand gold archival CDs had an advertised life expectancy of 100 or 300 years, depending on the manufacturer. Our yearly check of collections burned to disc since 1995 found that between 2% and 5% of the discs failed annually.
- Long-term tape storage, such as Advanced Intelligent Tape (AIT) or Linear Tape-Open (LTO), usually read only two prior generations of tape, so the tape drives are often upgraded every 10 years, making the tapes obsolete sooner than the anticipated life expectancy. We have AIT2 tapes and the tape drives, but the drives are difficult to connect and to use.
- External hard drives and raid arrays have also failed and caused data loss; in one case, 8TB of master images were lost. With some important collections, one-third of external drives failed during the first year.

Typically there is no warning that digital storage is about to fail, so knowing when to refresh or to migrate media becomes a guessing game. Change too early and money is wasted. Wait too long and there is the potential of data loss. These experiences show why our institution now uses the M-Disc as one of our archival copies of long-term collections.

Each institution may have different storage policies and environments. Not every situation will be the same. By considering the criteria above (the storage costs, the average lifespan of the media and the migration time frame), institutions can make a more informed choice about their archival digital storage environment.

General Terms

Infrastructure opportunities and challenges; Frameworks for digital preservation

Keywords

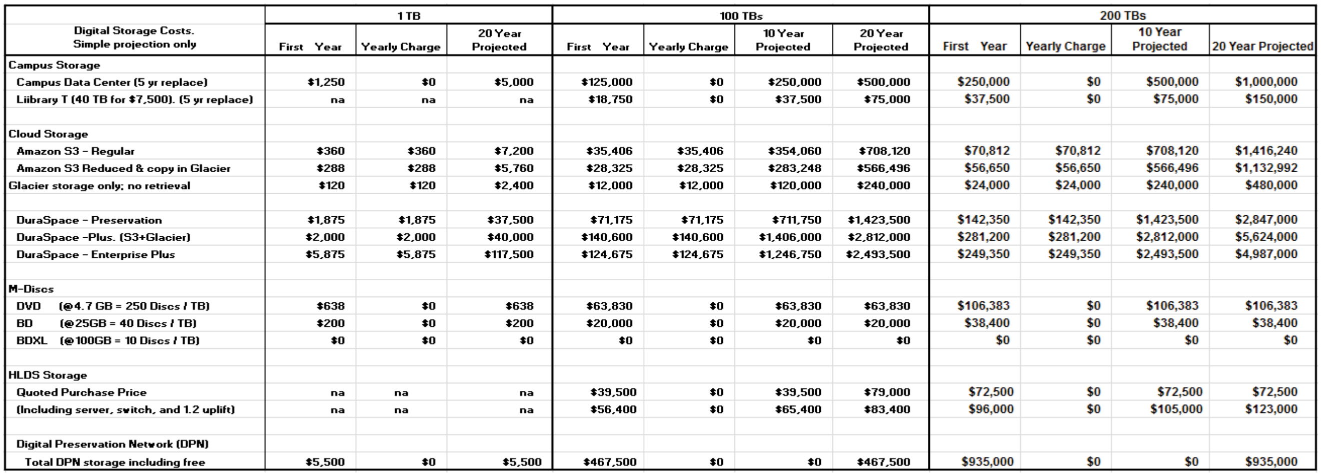
Digital Preservation; Digital Storage; Storage Costs.

REFERENCES

[1] Chapman, Stephen. Counting the Costs of Digital Preservation: Is Repository Storage Affordable? *Journal of Digital Information*. 4, 2 (February. 2004), n. pag. DOI= https://journals.tdl.org/jodi/index.php/jodi/article/view/100.

[2] DeRidder, Jody L. Considerations for Storage and Protection of Content. *ASERL Webinar: Library Publishing*. (February. 2012), 14. DOI=http://www.aserl.org/wp-content/uploads/2012/02/Store_protect.pdf.

[3] Wright, Richard, Ant Miller, and Matthew Addis. The significance of storage in the "cost of risk" of digital preservation. *International Journal of Digital Curation*. 4, 3 (2009), 104-122. DOI=http://www.ijdc.net/index.php/ijdc/article/view/138.



	1 TB			100 TBs				200 TBs			
Digital Storage Costs. Simple projection only	First Year	Yearly Charge	20 Year Projected	First Year	Yearly Charge	10 Year Projected	20 Year Projected	First Year	Yearly Charge	10 Year Projected	20 Year Projected
Campus Storage											
Campus Data Center (5 yr replace)	$1,250	$0	$5,000	$125,000	$0	$250,000	$500,000	$250,000	$0	$500,000	$1,000,000
Liibrary T (40 TB for $7,500). (5 yr replace)	na	na	na	$18,750	$0	$37,500	$75,000	$37,500	$0	$75,000	$150,000
Cloud Storage											
Amazon S3 - Regular	$360	$360	$7,200	$35,406	$35,406	$354,060	$708,120	$70,812	$70,812	$708,120	$1,416,240
Amazon S3 Reduced & copy in Glacier	$288	$288	$5,760	$28,325	$28,325	$283,248	$566,496	$56,650	$56,650	$566,496	$1,132,992
Glacier storage only; no retrieval	$120	$120	$2,400	$12,000	$12,000	$120,000	$240,000	$24,000	$24,000	$240,000	$480,000
DuraSpace - Preservation	$1,875	$1,875	$37,500	$71,175	$71,175	$711,750	$1,423,500	$142,350	$142,350	$1,423,500	$2,847,000
DuraSpace -Plus. (S3+Glacier)	$2,000	$2,000	$40,000	$140,600	$140,600	$1,406,000	$2,812,000	$281,200	$281,200	$2,812,000	$5,624,000
DuraSpace - Enterprise Plus	$5,875	$5,875	$117,500	$124,675	$124,675	$1,246,750	$2,493,500	$249,350	$249,350	$2,493,500	$4,987,000
M-Discs											
DVD (@4.7 GB = 250 Discs / TB)	$638	$0	$638	$63,830	$0	$63,830	$63,830	$106,383	$0	$106,383	$106,383
BD (@25GB = 40 Discs / TB)	$200	$0	$200	$20,000	$0	$20,000	$20,000	$38,400	$0	$38,400	$38,400
BDXL (@100GB = 10 Discs / TB)	$0	$0	$0	$0	$0	$0	$0	$0	$0	$0	$0
HLDS Storage											
Quoted Purchase Price	na	na	na	$39,500	$0	$39,500	$79,000	$72,500	$0	$72,500	$72,500
(Including server, switch, and 1.2 uplift)	na	na	na	$56,400	$0	$65,400	$83,400	$96,000	$0	$105,000	$123,000
Digital Preservation Network (DPN)											
Total DPN storage including free	$5,500	$0	$5,500	$467,500	$0	$0	$467,500	$935,000	$0	$0	$935,000

Figure 1: Estimated Costs of Digital Preservation Storage

(Re-)publication of Preserved, Interactive Content – Theresa Duncan CD-ROMs: Visionary Videogames for Girls

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

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

ABSTRACT

This poster presents implementation details, reception and lessons learned from a Cloud-based emulation project for world-wide interactive access to preserved CD-ROMs.

General Terms

Infrastructure opportunities and challenges; Frameworks for digital preservation

Keywords

Emulation, CDROM Preservation, Access, Case-Study

1. INTRODUCTION

A core mission of Rhizome, a born-digital art institution founded in 1996, is to make it possible to experience digital art on the Internet. A piece of software art that is part of a collection but cannot be accessed or circulated within the current conditions of digital communication, doesn't make much sense. Especially, software needs to run and be experienced, it cannot be substituted with representational media. Hence, making legacy software accessible for a (world-)wide audience is an important challenge, in particular if the items cannot be downloaded and executed locally, but require an outdated operating system and hardware to run.
Within the project *The Theresa Duncan CD-ROMs* Rhizome's goal was to re-enact three style-defining art game CD-ROMs from the 1990's on the web. Without users having to install any additional software or download an emulator or a disk image, the games should be played in their completeness on any modern browser. Rhizome launched a Kickstarter project [1] to make the three titles–*Chop Suey*, *Smarty* and *Zero Zero*–available to their audience free of charge.
This poster presents implementation details, usage statistics and user perception of a Cloud-based emulation experiment, using *Emulation as a Service* (EaaS)[2].

2. THERESA DUNCAN'S VIDEO GAMES FOR GIRLS

The Theresa Duncan CD-ROMs, published in between 1995 and 1997, are culturally important and pioneering "female games," but have been out of print for more than 15 years and remain inaccessible on contemporary computing hardware. These titles represent an important counterbalance in the "pink gaming" explosion of 1990's girls' CD-ROMs that were dominated by the template of the highly successful Barbie series, perpetuating a very traditional image of girlhood.

Figure 1: Chop Suey (Magnet Interactive, 1995, co-created with Monica Gesue) one of the three CD-ROMs re-published online. Lily and June Bugg embark on a strange, hallucinatory adventure through the small town of Cortland, Ohio.

2.1 Dramaturgy & Technical Requirements

None of the three games features fast-paced action or 60 Hz animation, like a typical 8-bit action game, but use the CD-ROM medium to present large amounts of animations, audio, and lots of surprises in the game's narration. None of the games is linear, there is no "progress" to be achieved, and no game status to be saved.
So there is no benefit of increasing the reaction time of the

[1] https://www.kickstarter.com/projects/710593842/theresa-duncan-cd-roms-visionary-videogames-for-gi

[2] http://eaas.uni-freiburg.de

software by running an emulator locally; the high volume of data never needs to be downloaded; since no states of the system need to be saved, there is also no need for accounts or write-back storage. Instead, the easiest possible access to the software could be established via a simple link, comparable to a YouTube video. EaaS runs emulators on remote computers and offers a ”window” via a web browser to interact with them. EaaS not only creates the possibility to run legacy computing environments and make them available on the web, but also allows the steward to control access to the objects. Users can experience the games in their full interactivity without having or being able to download the game data.

The CD-ROMs use a hand-drawn pixel graphic style that cannot be meaningfully represented with typical video codecs, like h264 or VP8, which are optimized for lens-based motives. The games' fine 1-pixel lines and patterns would result in halos and unpleasant artifacts. EaaS uses lossless (compressed) graphics to deliver pixel-perfect graphics to users. The rich soundtracks of the games use uncompressed 16 Bit 22.05 kHz audio, and feature long voice-narrated sequences and music. Continuity of the sound is much more important than audio quality. Additionally, most of the audio was created in home-recording settings, which adds to its charm and affect. For this project EaaS was configured to use OGG/Vorbis audio compression to stream sound, since it offers a balance between speed of encoding, quality and browser support.

Unlike online video streaming services, the output of an emulator's screen can not be buffered or pre-fetched, since it is dependent on user input. Hence, a short network distance between the emulator's computing node and the user accessing it is highly desirable. To achieve low network latency, we chose cloud computing services[3] to be able to allocate computing resources close to the user's location.

Different legacy operating systems were tested for the reenactment, the games themselves support as many as four different ones. Goals were: a quick start of the game after a user gets access to an emulator, system stability, reasonably authentic performance, and usability. Macintosh System 7.5 proved to be the best option, other systems being ruled out due to slow startup time, Quicktime video rendering problems or unstable mouse pointer movement.

3. TECHNICAL BACKEND

The technical setup for this project consists of three core components, the Rhizome web site, an EaaS gateway and a dynamic number of emulator compute nodes running in the cloud.

Rhizome's web site provides the front-end for users to select a game for emulation. It further implements a user queue and issues bulk requests to the EaaS gateway (e.g. three sessions for Chop Suey, five sessions Zero Zero). The EaaS gateway processes these requests by assigning free CPUs, ie. pre-allocated CPUs, to requested emulated environments and responds with one *iframe*-URL for every newly created emulation session.

In case of insufficient CPU resources, the EaaS gateway allocates a new cloud machine and returns less emulation sessions than requested. It is then up to the front-end to manage the waiting queue. If the user has to wait for a compute node to become ready, the front-end displays a waiting animation. If a single user's request was successful, the front-end embeds the iframe-URL, enabling interaction with the emulated environment, ie. playing one of the CD-ROM games. Once an emulation session is active, the EaaS gateway is in charge of session management. The session ends when the user leaves the specific emulator page (or closes the browser window or tab). The EaaS backend recognizes an expired session and releases all resources used, especially paid on-demand computing resources.

In coordination with the online publication of The Verge[4], the games were disseminated online and embedded into several online magazines, personal blogs and gaming sites, just like a regular youtube video. For each user one virtual CPU was assigned at the Google Compute Cloud's (US-central). During the peak phase, when the project was just disseminated and discussed on social media, 16 CPU machines were used, always preallocating 16 spare CPUs on top of the current demand to reduce potential waiting time. After the first big rush, smaller machines were allocated.

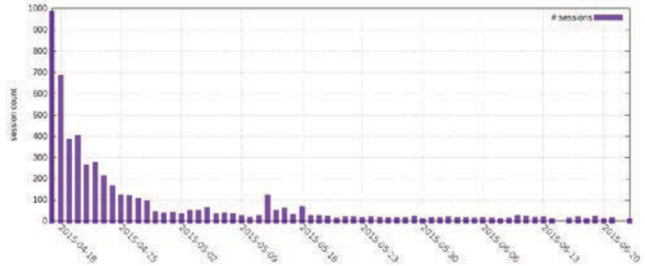


Figure 2: EaaS sessions per day.

4. PERCEPTION AND USAGE

From launch of the project April 17 to June 23, 4644 emulation sessions were served, from that 976 sessions during release day (cf. Fig. 2). During the launch phase, users mostly tried the games out very briefly. For the plateau phase, the usage pattern changed to less users that were more ”devoted” and played the games for up to two hours. The median session time was 99 seconds, with a wide variance between users. Top-20 users' session time was at least 109 minutes.

The online versions of the CD-ROMs have been discussed and embedded on *The Verge*, *itch.io*, *Artforum* and the *Huffington Post*, along with a few personal *alt.game* blogs. Complete, hour-long emulation play-throughs created by enthusiasts of all three titles appeared on YouTube. A *github* snippet of a simple `iframe` code was circulated on social media, enabling the games to be embedded into any website.[5]

Interestingly, none of the publications or fan productions were paying much attention to the technical form of delivery of the games, but rather indulged in cultural analysis and interpretation. This can be seen as sign of EaaS functioning reliably and transparently.

[3] Google Compute was chosen because of the very quick deployment process and the simple, predictable pricing structure.

[4] http://www.theverge.com/2015/4/17/8436439/theresa-duncan-chop-suey-cd-rom-preservation

[5] https://gist.github.com/despens/098823cd5b6c577fb5a5

Open Preservation Foundation Community Survey 2015

Ed Fay, Becky McGuinness, Carl Wilson
Open Preservation Foundation
c/o The British Library
Boston Spa, LS23 7BQ, UK
+44 (0) 1937 546013
ed@openpreservation.org
becky@openpreservation.org
carl@openpreservation.org

Nick Krabbenhoeft
CodedCulture
New York, NY 11377 USA
+1 907 409 7907
nick@codedculture.org

ABSTRACT

This poster will present the headline results from the Open Preservation Community Survey 2015, which surveyed over 130 institutions around the world to establish the current state of the art in digital preservation practice. The survey focused on technology adoption and real-world infrastructure and architectures, including demographics about the type and size of the responding institution. The responses include: staff roles and allocations; core digital preservation activities; content types accepted for long-term management; storage capacity and models; use of the cloud and consortial solutions; use of open source; repository and workflow systems; and tool adoption. The survey did not ask about policies or costs. In addition, comparisons are drawn with the PLANETS survey [1] from 2009 to show changes in requirements and practice over time. The published analysis and raw data will be forthcoming by the end of 2015.

General Terms

Infrastructure opportunities and challenges; Preservation strategies and workflows.

Keywords

Digital preservation practice; open source; technology adoption; continuous improvement.

1. ACKNOWLEDGMENTS

Our thanks to the 130 institutions that responded to the survey and gave their permission for their responses to be included in the data to be published openly. Thanks also to those who gave comments on drafts of the survey instrument and assisted with the recruitment of respondents: Trevor Owens, Angela Goethals, Neil Beagrie, Herve L'Hours, Jack Kneeshaw, Hersh Mann, Angela Dappert, Catherine Jones, Tom Cramer, Alastair Dunning, Zoltan Szatucsek, and anyone else we may have omitted to mention for which the cause is accident and the fault ours alone.

2. REFERENCES

[1] Jardine, L., Sinclair, P. 2009. PLANETS survey analysis report. http://www.planets-project.eu/docs/reports/planets-survey-analysis-report-dt11-d1.pdf

The Strategic Framework and the Mechanism of Rights Management of Long-term Preservation

Yin Gaolei
National Science Library, Chinese Academy of Sciences
National Science Library, Chinese Academy of Sciences, Beijing, 100190
+86-010-82626611-6107
yingaolei@mail.las.ac.cn

Zhao Yan
National Science Library, Chinese Academy of Sciences
National Science Library, Chinese Academy of Sciences, Beijing, 100190
+86-010-82626611-6129
hnzzu622@126.com

ABSTRACT

In this paper, we describe how the National Science Library, Chinese Academy of Sciences is building the Strategic Framework and the Mechanism of Rights Management for Long-term Preservation of digital resources.

General Terms

Frameworks for digital preservation

Keywords

Long-term Preservation, Strategic Framework, Mechanism of Rights Management

1. INTRODUCTION

In this paper, we describe how the National Science Library, Chinese Academy of Sciences is building the Strategic Framework and the Mechanism of Rights Management for Long-term Preservation of digital resources. Firstly, we introduce the strategic objective of long-term preservation on digital resources of our library according to its responsibility. Then we present our library's rights claim on long-term local preservation as well as the targeted resources. Finally, we elaborate how our library to establish institutional right management based on agreement and how to establish working procedures and mechanism of management and operation in accordance with industry standards.

2. Strategic Objective

As an institute providing information services for researchers and students within Chinese Academy of Sciences (hereinafter referred to as "CAS"), National Science Library, Chinese Academy of Sciences(hereinafter referred to as "NSL")also serves as a national information institute specialized in basic sciences, interdisciplinary fields, and strategic high-tech fields. In order to strategically safeguard scientific and technical information resources for CAS and even the whole country, it is NSL's obligations publicly and professionally to undertake the long-term preservation of digital resources as the internal and key work when building a complete guarantee and service mechanism of information resources, to research and solve the critical problems, and to develop reliable capability to long-term preserve main digital resources of which users in our country are in need.

3. Right Claim

NSL advocates the rights of long-term local preservation of the purchased digital resources:

- The long-term preservation right means that the library has reasonable archiving, processing, serving and cooperative depositing right of the purchased resources.
- The right of long-term preservation is an integral part of library's rights of purchased knowledge content.
- The long-term preservation right is a prerequisite to provide reliable utilization for the users and one of the important measures in support of digital resources' sales, promotion and application.
- The right of long-term preservation is an important cooperative basis for libraries to continue purchasing digital resources from the digital suppliers.
- The library fully acknowledges and protects the legal rights of publishers in the long-term preservation of digital resources.
- From the year of 2014, NSL has taken the authorization of the long-term preservation right as basic requirements and widespread appeals of procurement.

4. Target Resources

The targeted resources NSL is planning to long-term preserve include the following:

1) The scientific and technological Journal databases and conference proceeding databases in basic sciences, interdisciplinary fields, and strategic high-tech fields published by major domestic academic journal publishers and major international comprehensive publishers.
2) The scientific and technological journal databases and conference proceeding databases in basic sciences, interdisciplinary fields, and strategic high-tech fields published by major professional associations and societies and specialized publishers.
3) Important open access journals, open access conference proceedings, open access professional knowledge repositories and other academic resources in basic sciences, interdisciplinary fields, high tech fields, and comprehensive scientific and technological fields.

4) Important digital academic monographs in basic sciences, interdisciplinary fields, high-tech fields, and comprehensive scientific and technological fields.
5) Digital academic journals, conference proceedings and monographs home and abroad incomprehensive and other scientific and technological fields.
6) Other important knowledge resources, such as dissertations, academic reports, research archives, etc. in basic sciences, interdisciplinary fields, high tech fields, and comprehensive scientific and technological fields.

5. Right Management

NSL is committed to establish institutional management to ensure rights and obligations of all parties being fully acknowledged and executed in the process of resource acquisition, preservation, public services and so on:

5.1 Agreement Framework

NSL has established the framework of long-term preservation agreement, in order to fully protect the legal interests of the stakeholders. According to this framework, NSL will sign a legally-binding long-term preservation agreement(such as , NSL has signed Supplementary Agreement on Long-term Preservation of Licensed E-journals with Springer as well as Wiley. However, NSL also has cooperated with RSC on Long-term Preservation of Licensed E-journals just according the ELECTRONIC ACCESS LICENCE AGREEMENT which including Journal Archive terms) with specific resource provider during the cooperative process of long-term preservation, which is an integral part of the purchase agreement.

The framework is a contractual mechanism of management, which must be consistent with the public interests and legal requirements and fully balance the legal interests of the stakeholders. It includes the following:

1) The ranges and contents of preserved resources, authorized institutions, authorized users, etc. have been defined.
2) The legal rights of resource preservers and providers involved in the long-term preservation have been defined.
3) It clearly states the conditions of getting access to, obtaining and testing preserved data, right and interest requirements in data processing, disputes resolution, etc.

5.2 Trigger Event

Trigger event is defined in accordance with mainstream international agreements and the Portico preservation agreement. Only by the following trigger events, can long-term preservation emergency services be started and strictly authorized:

- Licensor No Longer in Business;
- Title No Longer Offered;
- Back Issues No Longer Available;
- Force Majeure;
- Failure of local access.

5.3 Public Service

For the authorized users: the definitions of users of long-term preservation emergency services are consistent with the ones defined in the purchase agreements.

For the subscribed contents: the contents to which users are allowed to get access by long-term preservation emergency services remain the same as the ones defined in the purchase agreement.

6. Mechanism of Management and Operation

6.1 Establish working procedures and mechanism of management &operation in accordance with industry standards.

1) Rights and interests in preservation: to initiate a system of legal rights and interests covering the entire process of long-term preservation; to work out legally-binding regulations and procedures which are practicable.
2) Archiving process: according to ISO 14721:2003, to perform integral management of the process including data acquisition, data preservation, data access, storage management, preservation management.
3) Preservation system: following the international standards relating to trustworthy digital repositories, this system can support the reliable running of the long-term preservation in an effective and economical way.
4) Backup and inheriting preservation: to provide multiple-level backup strategies and methods based on trustworthy requirements, ensuring the usability, validity, and time effectiveness.
5) Auditing and certificating: in light of ISO 16363 and other standards of trustworthy repositories around the world, a public inspecting mechanism of long-term digital documents preservation is to be built up, which is trustworthy and involves joint efforts.
6) Public service: guided by international reliable standards of the public service in long-term preservation, public service is going to be managed according to the most proper practices and the optimum mechanism in the field.

6.2 A Sustaining and Steady Input Mechanism

NSL started the long-term preservation as a special project in 2009. And it has evolved into an important strategy these years. Until now, the project has been implemented for four terms with sustaining and steady funds from Chinese Academy of Sciences, which can be continued in a sustainable way.

6.3 Professional Crew and Organization

A center for long-term preserving national digital resources of science and technology under Chinese Academy of Sciences has been built, and equipped with professional crew. We have a team for collection development to perform negotiation on rights and interests. We also have a technology team to build the technology systems and to manage data. In addition we have a service team to deal with and manage emergencies.

7. ACKNOWLEDGMENTS

This work was supported by Foundation of long-term preservation, Chinese Academy of Sciences.

8. REFERENCES

[1] Portico. Triggered Content [EB/OL].[2015-09-15].

http://www.portico.org/digital-preservation

What We Teach: An Assessment of Graduate-Level Digital Curation Syllabi

Carolyn Hank
University of Tennessee
420B Communications Bldg
Knoxville, TN 37996-0341
1-865-974-4049
chank@utk.edu

Noah Lasley
University of Tennessee
451 Communications Building
Knoxville, TN 37996-0341
1-423-309-5684
nlasley@vols.utk.edu

Xiaohua Zhu
University of Tennessee
438 Communications Building
Knoxville, TN 37996-0341
1-865-974-3141
xzhu12@utk.edu

Kylan Shireman
University of Tennessee
451 Communications Building
Knoxville, TN 37996-0341
1-865-974-2148
kshirema@vols.utk.edu

Charlene N. Kirkpatrick
University of Tennessee
451 Communications Building
Knoxville, TN 37996-0341
1-865-974-2148
pgq766@vols.utk.edu

ABSTRACT

This poster reports preliminary results from an intensive review of English-language syllabi for Master's level courses in digital curation, undertaken as part of a larger study looking at the convergence and divergence of reading assignments between digital library and digital curation curriculum.

General Terms

Training and education

Keywords

Graduate; post-graduate education; content analysis; digital preservation

1. INTRODUCTION

The past decade has seen a rise in digital curation-specific and related curriculum offerings at graduate degree programs in information, library and archival science. Ray (2009) provides an early summary of digital and data curation curriculum development primarily in English-language programs, including funder support for such initiatives, with recent updates and considerations for further development presented by Tibbo (2015). This poster reports select findings specific to digital curation syllabi from a descriptive content analysis study of syllabi for digital curation and digital library courses offered at Master's degree granting programs. The aim of this broader study is to identify the extent foundational courses in digital library and digital curation converge and diverge, as evidenced through stated course objectives, reading assignments and other syllabi characteristics. This study drew inspiration from the Digital Library (DL) Curriculum project's research identifying the core literature offered in digital library courses from graduate programs accredited by the American Library Association (ALA) (Pomerantz et al, 2006). The source listings for this study were expanded to also include programs listed to the Society of American Archivists' (SAA) Directory of Archival Education and the iSchools' Directory.

2. METHOD

The objective for the study results reported in this poster is the extent to which a core of frequently assigned digital curation readings exists. Further, it examines the diversity of reading assignments by format and content type, with particular emphasis on assignments from academic and scholarly journals. This is in response to the questions of whom and where: which authors are frequently assigned, and in what journals are they publishing?

2.1 Population and Sampling

An aggregate listing of the ALA, SAA, and the iSchools' respective directories was compiled, resulting in a list of 101 programs. These directory affiliations are not exclusive. A quarter of the programs (26%) are listed to two source listings; twelve (12%) are listed to all three sources. While syllabi are the data source for this study, the first step for gathering syllabi was a review of the 101 programs' respective websites. Review was limited to programs with website content available in English. As a result, eight programs were excluded from further consideration. For the remaining 93 programs, their websites were examined to locate recent course schedules, characterized as a list of courses arranged by academic term and with sufficient course details, including instructor, course meeting time and course number, to complete a web search for the respective course syllabi. Though most of the 93 programs made course catalogs available on their websites (71, or 76%), only 35, or 38%, made recent course schedules available for one or more academic years.

For these 35 programs, course schedules were reviewed to identify any courses offered in digital curation. In addition to searching by the key words, "digital curation," within the course title or description, other related key words were searched, including digital preservation, digital archiving and digital stewardship. Nearly half of the programs (16, or 46%) did not have course offerings with these or related key words.

For the remaining 19 programs that did, the courses identified were further assessed for eligibility to enhance homogeneity among the sampled syllabi. The focus was on foundational courses in digital curation, rather than courses in advanced digital curation topics. Courses that were deemed to provide more depth in a particular area, rather than breadth across digital curation as a whole, were excluded, such as digital forensic courses or courses dedicated to particular format or content type, such as moving images or research data. In regard to the latter, courses in data curation were excluded. Additionally, courses in data science were outside the scope of this study.

Additionally, as the course schedules reviewed represented two or more academic years, this contributed to duplicate course offerings. After removing for duplicates, 15 syllabi were identified and collected. While these sampling design decisions resulted in substantially fewer syllabi eligible for analysis, it did lend support in identifying unique syllabi rather than potentially similar or identical syllabi from the same programs.

2.2 Data Collection and Analysis

Two coding systems were created to capture syllabi attributes. The first collected information on the front matter contained within the syllabi, such as course delivery mode. The second collected information from the reading outlines of the syllabi, allowing for capture of each assigned reading whether indicated as required or optional. If no indication as to status, then it was assumed (and coded) to be required. A citation key was created to capture and manage each unique assigned reading as identified in the reading outlines and for coding all assigned readings for format and content type, authorship and publication channel. Once syllabi coding was complete, the data collected was reviewed and normalized. The citation key in particular demanded extensive cleanup due to the variation in how assigned readings are formatted in the syllabi and the degree of completeness of bibliographic data for citations provided. All citations were reviewed and normalized, with efforts made to complete missing essential bibliographic data. The syllabi front matter, reading outline and citation key data were preliminarily analyzed in Excel and then exported to SPSS for final analysis.

3. Select Results

Overall, the fifteen syllabi coded represented three academic periods: Fall 2012 (1); Winter or Spring 2014 (5); and Fall 2014 (9). Considering the source directories from which the program sampling frame was constructed (ALA, SAA and iSchools), seven syllabi (46%) were listed to all three directory sources.

3.1 Syllabi Front Matter

The 15 syllabi were assessed for degree of completeness. Eleven (73%) are complete, containing a reading outline. Of these, nearly (10 out of 11) are arranged by topical theme. Front matter was also examined to identify presence of typical syllabus components, such as: course description (87%), course objectives (87%), course topics (13%), instructional method (80%), and assignments (80%). Front matter was also examined to identify presence of required and/or optional textbooks. A majority did not indicate required (67%) or optional (80%) textbooks.

3.2 Syllabi Reading Outline

From the eleven digital curation syllabi containing reading outlines, all reading assignments were collected and coded. After extensive, iterative review to remove for duplicates, a total of 729 unique citations were identified. "Assignments" is preferred for referring to these "readings" to reflect that not all citations are text-based. Ultimately, 24 categories were derived, representing more formal and traditional sources, such as books and journals, to more informal as well as transformative sources, such as white papers and grey literature to blogs and web-based videos. Refereed journal articles are the largest grouping, accounting for 34% of all citations. Refereed, and to some degree non-refereed journal articles and trade articles, are typical indicators of impact in bibliometric studies. While, in combination, these three serial types account for 335 (43%) of all unique citations, non-serial publications are also well-represented, specifically organizational and research project publications. In combination, these account for 184 (25%) of all unique citations.

For the 247 unique refereed journal articles assigned, 72 refereed journals were represented among these 247 citations. A majority (53, or 74%) contributed only one article, with four journals publishing two articles each. The remaining 186 articles were published in 15 journals, with three (*International Journal of Digital Curation, American Archivist and Archivaria*) publishing a combined 155 articles, representing nearly half (47%) of all the unique articles assigned.

Authors of academic journals (refereed and non-refereed) were examined. For these combined 280 articles (247 refereed and 33 non-refereed), 171 were singularly authored, and 109 had two or more authors. An author index was compiled of all attributed authors, single or collaborative, resulting in a listing of 489 unique authors. For the 489 authors listed, 318 (65%) were listed once, contributing one article, either single-authored (n=99) or co-authored (n=219). For the 280 refereed and non-refereed articles, 26 are the work of just four authors. Hence, these four authors produced nearly one out of 10 (9%) of all articles. The most frequent authors, characterized as contributing four or more articles, are: Duranti, L.; MacNeil, H.; Bearman, D.; Conway, P.; Hedstrom, M.; Rosenthal; D.; and Ross, S.

Returning to the 729 unique assignments, a total of 927 required and optional assignments were made in the 11 reading outlines, for an average of 84 readings per syllabus. As the study is interested in frequency of assignments among syllabi, rather than within syllabi, assignments were reviewed to identify any duplicate assignments; that is, the same assignment being listed two or more times within the same reading outline. This reduced the number of assignments – with only one per syllabus considered – from 927 to 868 total assignments. For these, 647 (76%) were required and 221 (26%) optional. The number of syllabi to which these total assignments (required and optional) are made clearly shows a long tail distribution, with a near majority (846, or 98%) assigned only once among the 11 reading outlines. Considered separately by assignment status, 627 (97%) required assignments and 219 (99%) optional assignments were only assigned once.

4. REFERENCES

[1] Ray, J. 2009. Sharks, digital curation and the education of information professionals. *Mus. Mgmt. and Curatorship.* 24, 4 (Dec. 2009), 357-368. DOI=http://doi:10.1080/09647770903314720

[2] Tibbo, H.R. 2015. Digital curation training and education: From digitalization to graduate curricula to MOOCs. *Intl. J. Digital Curation.* 10, 1 (Feb. 2015), 144-154. DOI=http://doi:10.2218/ijdc.v10i2.345

[3] Pomerantz, J.P., Oh, S., Yang, S., Fox, E.A., and Wildemuth, B.M. 2006. The core: Digital Library Education in Library and Information Science Programs. *D-Lib Mag.* 12, 11 (Nov, 2006). DOI=http://doi:10.1045/november2006-pomerantz

Software Reuse, Repurposing and Reproducibility

Catherine Jones
STFC
Harwell Oxford
Didcot
44 (0)1235 445402
Catherine.jones@stfc.ac.uk

Brian Matthews
STFC
Harwell Oxford
Didcot
44 (0)1235 446648
Brian.matthews@stfc.ac.uk

Ian Gent
St Andrews University
North Haugh
St Andrews
+44 (0)1334 46 3247
Ian.gent@st-andrews.ac.uk

ABSTRACT

Software underpins the academic research process, regardless of discipline. With the increased focus on the long term value of data and other research outputs, then more attention needs to be paid to how software used in these processes is both identified and preserved for the long term as much data is meaningless without the related software. In this poster we describe the aims, objectives and current results of the Jisc funded project Software Reuse, Repurposing and Reproducibility (Software RRR). This poster discusses the issues around persistently identifying software, makes some recommendations for good practice, and discusses the relationship between identifying source code and a playable version of this software.

General Terms

Preservation strategies and workflows

Keywords

Software preservation

1. INTRODUCTION

Software underpins the academic research process, regardless of discipline. Software is written to be run, and while programmers might strive for elegance or beauty in the code, the overwhelming point of software is to execute it. To be able to understand and use/reuse and preserve data then the software code which generated, analysed or presented the data will need to be retained and executed. A starting point is the persistent identification of software to maintain the integrity of software as an item over time. This is an emerging area and services such as Zenodo (https://zenodo.org/) are enabling developers to persistently identify code.

Software is a composite artefact and may have different components bundled together. This can be seen by the following definition:

> *"Computer software includes computer programs, libraries and their associated documentation. The word software is also sometimes used in a more narrow sense, meaning application software only."(* https://en.wikipedia.org/wiki/Software Retrieved 12/6/2015)

Consequently, software cannot be treated in a similar manner to other digital artefacts (for example documents, media content or data) and needs separate consideration for preservation. Further, if the software is to remain reproducible and reusable, additional consideration needs to be taken to maintain its correct execution behaviour.

1.1 Aims of the software RRR project

The Software RRR project is investigating the persistent identification of software and how links can be made to runnable versions of software enabling preservation of functionality. The project builds on the Recomputation project (http://recomputation.org) [1] and earlier work on a framework for software preservation [2],[3].

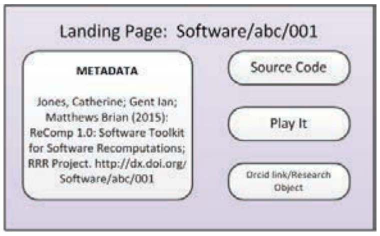

Landing Page: Software/abc/001
METADATA
Jones, Catherine; Gent Ian; Matthews Brian (2015): ReComp 1.0: Software Toolkit for Software Recomputations; RRR Project. http://dx.doi.org/Software/abc/001
Source Code
Play It
Orcid link/Research Object

Figure 1 Landing Page

The figure above represents the vision of the project which is encapsulated in a landing page for a persistently identified software object with effective metadata, links to the downloads, including source code and a runnable version, together with hooks to other entities in the wider context such as Orcid, data and publications. Thus a user can: uniquely identify software released in a particular context (via software citation); access landing pages which give additional metadata to describe the software; access a runnable version of the software replicating its original behavior; and download packages with sufficient information to allow its reconstruction locally.

To realise this vision, we need to provide consistent guidelines for software identification together with local metadata and a virtualized platform for replay and recomputation. In the rest of this short paper, we concentrate on issues of persistently identifying software.

2. ISSUES IN PERSISTENT IDENTIFICATION

2.1 What is being identified?

A key issue is what exactly is being identified, as described in the previous section software is a complex object and may include one or more of: source code, executable version, packaged version, additional items such as included libraries and documentation. Further, software typically is an evolving artefact, with different expressions being made available through a software release cycle, reflecting the changes in functionality and computing environment which software undergoes over time.

We use a four level model of software to describe the different expressions a software system has over its release cycle. This model enables better understanding of what might need to be persistently identified.

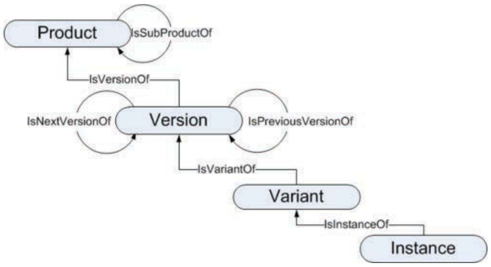


Figure 2 Levels of Software

- **Product:** The whole top-level conceptual entity encompassing the whole lifecycle of the software, and is how the system may be commonly or informally referred to
- **Version:** is an expression of the product which provides a single coherent presentation of the product with a well-defined functionality and behaviour and usually in how it interacts with the computing environment.
- **Variant** is a version adapted to a specific operating environment
- **Instance** is an actual deployment of a software product which is to be found on a particular environment or machine.

A particular software citation will typically refer to a particular expression of the software which is used in a particular context, thus the same expression should be used to validate the results.

2.2 Environment

The environment that the software was built and operates in is a vital part of ensuring software is not just preserved but remains runnable. Metadata supplied with the software expression should describe its target environment. This is a complex area and has not been addressed in this phase of the project.

2.3 Metadata

Metadata plays an important role in the discovery, access, management & preservation of software, and thus we need to consider the appropriate metadata to provide. We address the use of DataCite discovery metadata to describe software in the next section.

3. DATACITE METADATA

DataCite (www.datacite.org) issues Digital Object Identifiers for data and other research artefacts. While it is not the only persistent identification system available, its wide adoption means it is an important source for identification of software, and consequently we concentrate on how to adapt Datacite DOIs for the citation of software expressions.

Datacite provides set of metadata elements to characterise digital objects for search and discovery [4]. The DataCite elements have been analysed to propose an appropriate profile for describing software. The approach taken is not to prescribe the content of any specific element but to describe the importance and enable the potential user to establish the correct answer for theirown situation. This poster will discuss how some key elements are used in the context of identifying software.

3.1 Creator

This element identifies the people responsible for the software. However this may not be straightforward to ascertain as software has a long life-span and may be worked on by many people. The point during the development cycle that the first DOI is given may also affect those identified as creators.

3.2 Title

The title of the resource is a mandatory field and can contain significant information. In a software context, there are some specific issues. If it a piece of software written by a single person for a specific project does it actually have a name? Is the official name different from the common name? What effect is versioning or branching of code going to have on the name? Will the name used be unique enough for it to be found and distinguished from other search results?

3.3 ResourceType

There is a resource type of Software, but this is a rather wide category and at present there aren't many formal suggestions for how this might be broken down further. This is an area with potential for further work.

3.4 Description

This field is designed to enable the addition of further information to assist in the understanding of the object being identified. Currently the two subtypes being used for software are Abstract and Other. These do not encourage the use of this field for technical information that may be needed to understand the object and a new subtype with a more descriptive label may be of assistance

4. FURTHER WORK

The first phase of this project has been concerned with persistent identification. The next phase is concerned with how software may be captured in such a way to ensure it remains runnable, thus preserving the performance. Being able to link the different software artefacts together in a fixed complex object will enable the long term preservation of software

ACKNOWLEDGMENTS

This work has been funded by Jisc in the Research@Risk scheme. We are grateful to our colleagues Tom Griffin, Simon Dobson and John McDermott for their comments and advice.

REFERENCES

[1] Arabas, S. et. al. 2014, Case Studies and Challenges in Reproducibility in the Computational Sciences. 1st Summer School on Experimental Methodology in Computational Science Research, St Andrews, August 4-8, 2014. arXiv:1408.2123

[2] Matthews, B., Shaon, A., Bicarregui, J., Jones, C., Woodcock, J., and Conway, E. 2009. Towards a Methodology for Software Preservation. In 6th International Conference on Preservation of Digital Objects (iPres 2009), San Francisco, USA, 5-6 Oct 2009.

[3] Matthews, B., Shaon, A., Bicarregui, J., and Jones, C. 2010. A Framework for Software Preservation. International Journal of Digital Curation 5, no. 1.

[4] Datacite Metadata Working Group, 2015. DataCite Metadata Schema for the Publication and Citation of Research Data. Version 3.1, June 2015, doi:10.5438/001

Minimal Effort Ingest

Bolette Ammitzbøll Jurik
State and University Library
Victor Albecks Vej 1
DK-8000 Aarhus C
Denmark
+45 8946 2320
baj@statsbiblioteket.dk

Asger Askov Blekinge
State and University Library
Victor Albecks Vej 1
DK-8000 Aarhus C
Denmark
+45 8946 2100
abr@statsbiblioteket.dk

Kåre Fiedler Christiansen
State and University Library
Victor Albecks Vej 1
DK-8000 Aarhus C
Denmark
+45 8946 2036
kfc@statsbiblioteket.dk

ABSTRACT

In this poster we present the concept of *Minimal Effort Ingest* into a digital repository and discuss benefits and disadvantages of this approach.

General Terms

Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice.

Keywords

Digital Preservation, Digital Repositories, Minimal Effort Ingest, Ingest Workflow, Quality Assurance, OAIS

1. MINIMAL EFFORT INGEST

An expensive part of ingesting digital collections into digital repositories is the quality assurance (QA) phase. Traditionally, data and metadata are quality assured before ingest, to ensure that only data which complies with the repository data formatting and documentation standards is preserved. In Minimal Effort Ingest, we postpone the QA of data and metadata until after the data has been ingested and even further, if resources are not available. This approach makes it possible to secure the incoming data quickly.

There are benefits and disadvantages to this approach, as detailed below. At the State and University Library, Denmark, we have implemented Minimal Effort Ingest as the workflow for our *Newspaper Digitization Project* [4]. About 30 million newspaper pages are being scanned, and we receive about 50,000 scanned pages per day. We have built a workflow which first ingests the data and metadata into our repository and then performs QA on the ingested data. If a delivery is found to be invalid, a new scanning is requested. When the new delivery is received and approved, the old delivery is purged from the system.

It has proven easy to continually add additional checks to the QA, and to run these checks on both the new deliveries and the already approved content.

2. OAIS COMPLIANCE

It has long been standard to establish trustworthiness of a digital repository by a more or less strict compliance with the *Open Archival Information System (OAIS)* Reference Model [2].

In the OAIS model a *Submission Information Package (SIP)* is received into temporary storage, where QA is performed, then an *Archival Information Package (AIP)* which complies with the archive's data formatting and documentation standards is generated, and *Archival Storage* is updated.

In the Minimal Effort Ingest model the SIP is transformed into a minimal AIP and ingested directly into *Archival Storage*. QA is performed from the *Data Management Functional Entity* on data in *Archival Storage*. That means we have moved the QA step from the *Ingest Functional Entity*, where it is performed on SIPs, to the *Data Management Functional Entity*, where it is performed on the minimal AIPs.

Ingesting the SIPs into *Archival Storage* directly as described above appears to be in contradiction with the OAIS reference model. The QA is however still performed, and we thus claim that a repository implementing the Minimal Effort Ingest model will be, content- and preservation-wise, *eventually consistent* with a repository implemented in strict compliance with the OAIS model.

The State and University Library, Denmark has incorporated Minimal Effort Ingest into both it's *Digital Preservation Policy* [5] and *Strategy* [6]:

> "As soon as possible after a collection has been received, all data and metadata are ingested into the library's Repository to preserve the functionality of the digital collection. Once a collection has been ingested into the Repository, a number of preservation actions can be carried out. The owner of the collections and the system owner coordinate the activities."[5]

We audit the State and University Library, Denmark as a trustworthy digital repository using the ISO 16363 Audit and Certification of Trustworthy Digital Repositories Standard [3]. While this standard uses the common conceptual framework provided by OAIS, it does not require strict compliance with OAIS.

3. BENEFITS

Ingesting content early into the repository has a number of advantages.

3.1 Preserving as Early as Possible

By adding the content to the repository as early as possible, we ensure that the content is preserved, at the very least in it's binary representation.

3.2 A Consistent Platform

By ingesting the data and metadata into the repository system, we have a consistent platform for doing QA and normalization.

Instead of developing tools specific to the ingest workflow for a given collection, we create tools that work on the repository. This gives us a unified platform for the development process, and it also makes it easier to reuse the tools for different collections.

3.3 Repository Tools instead of Ingest Tools

QA and normalization tools can be used in other phases of the information flow than ingest. By making the tools into repository tools, we can run the tools whenever it is relevant.

This also ensures that any QA actions are performed on the same data we preserve. This is in contrast with an OAIS ingest workflow, where content conceivably might change in the interval between the QA step and the actual ingest step.

We can also update the QA tools, and rerun them on the collections, whether they are recently ingested or approved a long time ago.

3.4 Recording Preservation Events

Since all preservation actions are performed on content within the repository, it becomes natural to save information about the actions as metadata in the repository. In the newspaper digitization project [4], we use PREMIS [1] to store this metadata as preservation events.

3.5 Empowering Repository Managers

Since all tools now work on the repository content, it is much easier to empower repository managers to work with the digital preservation tools without involving IT resources.

In that way repository managers without special IT background can take responsibility for preservation actions.

4. DISADVANTAGES

Minimal Effort Ingest does have drawbacks.

4.1 Normalization

When working with normalization in an ingest workflow, it may result in having both pre-normalization and post-normalization copies of content in the repository. This requires, depending on policy, either twice the space, or a method for cleaning up in the repository.

4.2 Content Failing QA

If content is not approved by the QA process, it may be necessary to either delete content or replace content with a new version from the content provider if possible. This can be a problem, since repositories often have policies that content should never or rarely be deleted.

4.3 Malicious content

By moving QA process from the ingest phase to a process within the repository, we risk ingesting content that has not been analysed or filtered for malicious content.

This could lead to vulnerabilities, if the content is accessed before such a check can be made, or if the analysis software itself is vulnerable. Moving the QA process from the *Ingest Functional Entity* to the *Data Management Functional Entity* can be seen as an increased security risk. Extra care should be taken that malicious content in the repository cannot compromise the security of the repository.

5. CONCLUSION

All things considered, performing preservation actions post-ingest on the repository content, rather than during ingest provides benefits in both development effort and preservation liability.

6. ACKNOWLEDGMENTS

Thanks to the National Library Division and the IT Preservation Section at the State and University Library, Denmark, who took part in formulating and implementing the Minimal Effort Ingest approach in general and particularly for the newspaper digitization project. A special thanks go to Gry Vindelev Elstrøm, Knud Åge Hansen and Jette G. Junge. You know why!

7. REFERENCES

[1] Library of Congress. `http://www.loc.gov/standards/premis`, 2015. Accessed June, 2015.

[2] Space Data and Information Transfer Systems. *ISO 14721:2012 Open Archival Information System (OAIS) - Reference Model*. The International Organization of Standardization, 2012.

[3] Space Data and Information Transfer Systems. *ISO 16363:2012 Audit and Certification of Trustworthy Digital Repositories*. The International Organization of Standardization, 2012.

[4] `http://en.statsbiblioteket.dk/national-library-division/newspaper-digitisation/newspaper-digitization`, 2015. Accessed June, 2015.

[5] `http://en.statsbiblioteket.dk/about-the-library/DigitalPreservationPolicy_2014.pdf`, 2015. Accessed June, 2015.

[6] `http://en.statsbiblioteket.dk/about-the-library/DigitalPreservationStrategy_v3.pdf`, 2015. Accessed June, 2015.

Modeling Tweets in Compliance with the Portland Common Data Model

Martin Klein
University of California Los Angeles
Research Library
Los Angeles, CA USA
martinklein@library.ucla.edu

Kevin S. Clarke
University of California Los Angeles
Research Library
Los Angeles, CA USA
ksclarke@library.ucla.edu

ABSTRACT

The ingest of non-traditional digital library collections into a linked data-based institutional repository for archival and presentation purposes is challenging on many levels. We propose a model for Twitter data that is compliant with the Portland Common Data Model. Based on this model, we can derive a linked data serialization and perform the ingest into our preservation repository.

General Terms

Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice.

Keywords

Twitter, Portland Common Data Model, Fedora

1. INTRODUCTION

Digital library programs increasingly face the challenge of incorporating non-traditional collections into their preservation and presentation workflow. Examples of such collections for the UCLA library are video and text (transcripts, closed captions, on-screen text) from daily captured TV broadcast news, crawled and archived web pages relevant to particular topics, and social media content such as tweets, which are also collected on a per-topic basis.

Within UCLA's International Digitizing Ephemera Project[1], the Research Library is developing such non-traditional collections around the Egyptian Revolution in 2009 and the Iranian Green Movement in 2011. Aside from thousands of digital images, cell phone videos, and scanned flyers, the collections also contain social media content. In particular, the library has a dataset that consists of more than $400,000$ tweets from about $50,000$ distinct users on the topic of the Egyptian Revolution. These tweets are special in the sense that they all were sent from within a 200-mile radius surrounding the capitol city of Cairo and so potentially reflect the voices of activists on the ground rather than trained journalists from international media channels. It is our intention to incorporate these tweets into UCLA's library preservation and presentation framework. The underlying institutional repository is Fedora and the library is in the process of transitioning to Fedora version 4, which is based on the Linked Data Platform[2]. For collecting tweets we are using the open source software Social Feed Manager[3]. The tool obtains tweets from the Twitter API[4] in JSON format. The Portland Common Data Model (PCDM)[5] has recently gained a lot of traction in the community as a data model to describe resources. This description provides the basis for the (RDF) serialization and hence conveniently bridges the gap between an arbitrary resource (a tweet) and the ingest into an institutional repository (Fedora 4).

The contribution of this paper is a first approach of modeling tweets in compliance with the PCDM. We describe our model, the characteristics of its main components, and all their relationships (in an RDF sense). We are reporting on a work in progress and hence are actively seeking feedback from the community to help stabilize this model.

2. THE PORTLAND COMMON DATA MODEL

For a better understanding of the here presented model, we briefly summarize the three for us relevant components of the PCDM. In the PCDM, intellectual entities (works, digital objects) are modeled as objects. An object can have descriptive and access metadata associated with it, it can contain files, and even other objects. A group of resources (objects) is modeled as a collection. Collections can also have descriptive and access metadata and it has a link to all objects it aggregates. Objects and collections are per se a unordered sets but for use cases where the order matters, a proxy class can be used that establishes order via links and proper IANA relation types such as first, last, next, and previous. The bitstream (sequence of binary data) of a resource is modeled as a file. A file can be described by accompanying metadata such as size, content type, and provenance information.

[1] http://digital.library.ucla.edu/dep/

[2] http://www.w3.org/TR/ldp/

[3] http://social-feed-manager.readthedocs.org/

[4] https://dev.twitter.com/overview/api

[5] https://github.com/duraspace/pcdm/wiki

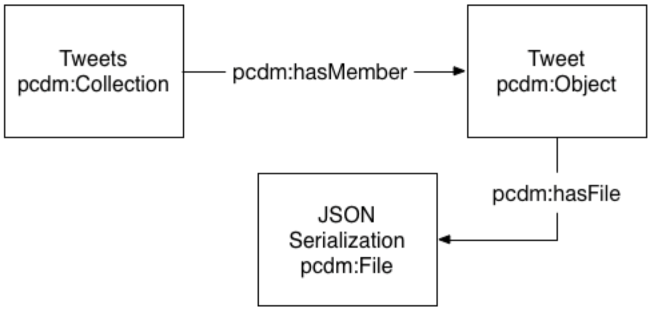


Figure 1: Collection, object, and file in the PCDM model

3. MODELING TWEETS

The diagram in Figure 1 depicts a high-level overview of our model. As we are capturing tweets by topic (a natural catastrophe, a political event) or from individual user accounts (UCLA athletics, a student organization), it makes intuitive sense that each tweet belongs to a collection, the high-level component defined in the PCDM. Each tweet is modeled as a PCDM object. Since these objects are member components of a PCDM collection, the collection links to each of them with the *hasMember* relation type. A tweet in JSON format comprises of a number of key/value pairs with notable examples being ID, text, created_at, and screen_name holding values for the tweet's unique numerical identifier, its textual content, the datetime it was sent, and the user name of its creator (the string following the @ character), respectively. One option for the PCDM would have been to deconstruct all or the for our use cases most relevant key/value pairs and model each of them individually as relationships of the tweet object. However, after consulting with the PCDM community, we decided against this approach and chose instead to model the JSON representation as a file which is linked to from the tweet object with the relation type `hasFile` as seen in 1. The main advantages of this approach is the retained simplicity and flexibility of the model. The simplicity comes from saving what otherwise would be several dozen links from the tweet object to the tweet's ID, text, creation date time, etc. and the flexibility is gained as different use cases can now individually chose which key/value pairs from the JSON serialization to process, for example in a Solr index to facilitate search.

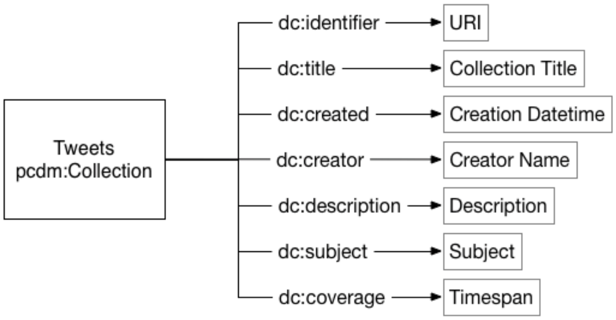


Figure 2: Link relations of the tweets collection

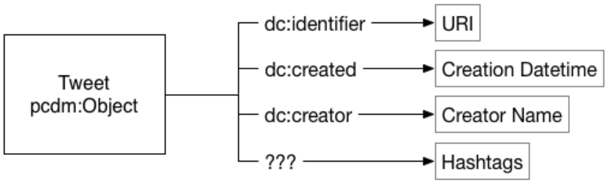


Figure 3: Link relations of the tweet object

All components in our model have associated descriptive and/or technical metadata, modeled as links with proper relation types. Once the RDF serialization of this model is ingested into Fedora, these data points can, for example, be queried via a SPARQL endpoint. The relations of the tweets collection are shown in Figure 2. Our model contains basic metadata elements for the collection-level such as the collection's title, URI as a unique identifier, subjects, and the timespan encompassing all component tweets. Figure 3 shows that our tweet object has four such links which all reference information directly derived from the tweet itself: its URI, creation datetime, the creator's name, and containing hashtags. This introduces a certain level of redundancy as the data also exists in the JSON file and will from there be indexed in Solr. However, it also enables us to process tweets on the RDF-level, for example, extract all tweets from particular users that contains certain hashtags. The links from the JSON file in our model are depicted in Figure 4. These links point to typical technical metadata of the file itself: its size, mime type, and hash value.

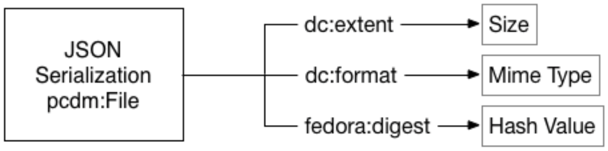


Figure 4: Link relations of the JSON file

Our model is not yet complete. For example, we have not identified a suitable relation type for the link between our tweet object and a hashtag that is contained in the tweet (as seen in Figure 3). Further, we have not yet sufficiently addressed the notion of access-level metadata but we are closely following the community discussion around the WebAccessControl system[6] and will adopt the emerging standard in due time. Also, a detailed discussion of the RDF-based linked data serialization of the model falls outside the scope of this paper.

4. SUMMARY

We introduce a model for tweets in compliance with the Portland Common Data Model in order to facilitate the ingest of such data into institutional repositories like Fedora 4. This model is still a moving target and we are actively seeking feedback on the here presented work. With the the ongoing community discussion and with feedback from the community, including the iPres audience, we are hopeful that we can derive a stable model for tweets in compliance with the PCDM

[6] `http://www.w3.org/wiki/WebAccessControl`

Mind the Gap. Bridging Digital Libraries & Archives

Mark Leggott
Islandora Foundation/ UPEI
550 University Ave.
Charlottetown, PE
1-902-314-7507
mleggott@islandora.ca

Erin Tripp
discoverygarden inc.
118 Sydney St
Charlottetown, PE
1-506-442-2520
erin@discoverygarden.ca

ABSTRACT

The overarching goal of digital libraries and archives are similar - to foster the preservation of cultural assets, published or otherwise. Despite this common goal, the different tools, terminology, and approaches of libraries and archives can impact, sometimes prevent collaboration. This presentation will highlight features of the open source Islandora framework that can help bridge the gap between digital libraries and archives, leveraging the strengths of both fields of expertise. The resulting rich ecosystem can provide an empowering approach to building user-centric collections that also achieve the goals of long term preservation.

This presentation will use examples from the Islandora Open Source Software Community and Framework to demonstrate the strides made to bridge the gap.

General Terms

Institutional opportunities and challenges; Frameworks for digital preservation; Innovative practice.

Keywords

Digital libraries, digital archives, open source, collaboration, Islandora.

1. INTRODUCTION

The overarching goal of digital libraries and archives are similar - to foster the preservation of cultural assets, published or otherwise. Despite this common goal, the different tools, terminology, and approaches of libraries and archives can prevent librarians and archivists from collaborating. This presentation will highlight features of the open source Islandora framework that can help bridge the gap between digital libraries and archives, leveraging the strengths of both fields of expertise. The resulting rich ecosystem can provide an empowering approach to building user-centric collections that also achieve the goals of long term preservation.

2. CONTENT

The OAIS reference model creates a common thread for discourse, especially with the increasing adoption of the model in both communities. The model will be discussed in relation to its application for digital libraries and archives, including non-traditional areas such as research data management. Also, new tools are emerging to bring digital libraries and archives even closer together. The three tools highlighted in this presentation will be the 1) the Manuscript Solution Pack (facilitating the viewing of a high resolution image of a manuscript, TEI and EAD description), 2) Drexel EAD modules (facilitates ingest and display of EAD and child object), and 3) Archidora (an integration between Islandora and Archivematica). Additional preservation features of the environment will be discussed.

3. CONCLUSION

The presentation will encourage leveraging existing knowledge bases and ecosystems for libraries and archives, working with the strengths of each, including standards, policies, workflows, and applications. Doing so will spur new developments that benefit both digital libraries and archives.

A National Preservation Solution for Cultural Heritage

Juha Lehtonen, Heikki Helin, Kimmo Koivunen, Kuisma Lehtonen, Mikko Tiainen
CSC – IT Center for Science
P.O. Box 405 (Keilaranta 14)
FI-02101 Espoo, Finland
+358 9 457 2001
{juha.lehtonen, heikki.helin, kimmo.koivunen, kuisma.lehtonen, mikko.tiainen}@csc.fi

ABSTRACT

We present the status of digital preservation at the National Digital Library (NDL) of Finland. The NDL has created a nationally unified structure for contents and services ensuring the effective and high-quality management, dissemination, and preservation of digital cultural heritage.

General Terms

Institutional opportunities and challenges; Innovative practice.

Keywords

Digital preservation, open source software, hardware

1. INTRODUCTION

The National Digital Library of Finland (NDL) is an entity within the remit of the Ministry of Education and Culture within the Finnish Government. The NDL ensures the preservation of digital cultural content, providing access to and compatibility of content, designing a cost-effective digital preservation solution, promoting the cooperation between the national libraries, archives and museums (partner organizations), and building better services with open cooperation and expansion to include a large range, and amount, of content.

Almost all memory organizations under the Ministry of Education and Culture of Finland are obligated by legislation to preserve cultural heritage. A major share of content owned and administered by partner organizations consists of digitized documents and photographs, but the volume of born-digital content is expanding quickly. Given of the diversity of the partner organizations, the digital content to be preserved makes up a very heterogeneous landscape. Based on extensive surveys conducted among partner organizations, we roughly estimate that digital information stored to our digital preservation solution by 2020 will consists of more than 2400 million objects requiring more than 12 petabytes of storage space not including necessary replication[1].

The NDL's digital preservation solution was taken into production during 2014 with about a half of a petabyte capacity. Although current capacity is somewhat moderate, our architecture is built to accommodate the increased volume and diversification of content and organizations, as well as the possible development into a storage system for the preservation of research data. In the spring 2015, the NDL's digital preservation solution was awarded the ISO 27001 information security certification, and we are planning for future auditing with preservation related standards, such as DSA (Data Seal of Approval) and ISO 16363.

The key activity in our solution to tackle preservation challenges is actively maintain a standard portfolio, which defines the standards to be used in the NDL. The national standardization ensures the functionality of the composite system, which requires semantic commensurability of metadata by partner organizations. All NDL specifications are produced in a close cooperation with partner organizations[2].

The standard portfolio, however, does not give detailed instructions for implementation or application of standards, but those are produced separately. The NDL METS profile defines a unified structure for Submission Information Packages (SIPs) and Dissemination Information Packages (DIPs). With a common digital preservation service for diverse organizations, the unified structure of information packages enables efficient administration of the information on the long term and also enables semantically commensurable information content. Having a common and mutually agreed format for both SIPs and DIPs helps partner organizations to build their own systems in a sustainable way.

Further, the NDL has specified a closed set of file formats that are accepted to our digital preservation service, with requirements of mandatory technical metadata elements for each content type. Currently, our digital preservation service supports two kinds of file format categories: Recommended and acceptable for transfer. Recommended file formats are such that the NDL considers to be usable for a long time, whereas acceptable for transfer are formats in which a significant amount of content is currently stored within partner organizations.

2. DIGITAL PRESERVATION SOLUTION

The software layer of the preservation solution implements the technical side of the digital preservation services by following the OAIS reference model. The software architecture consists of front-end (services for the partner organizations and for the system administrator), back-end (functionality and coordination services for task execution), and services for the storage.

Our digital preservation solution uses a highly modular microservice structure, which means that the architecture is divided into small, highly independent components. Such a component may be a 3rd party open source or an in-house software, which takes care of a certain part of the implementation. Also, we employ the idempotence property for the microservices, which means that a task in a workflow can be run several times so that the repetition does not affect to the final result nor functionality of the preservation solution. In other words, we are able to repeat the interrupted microservices, and skip the succeeded ones without clean-ups. Modularity is also necessary for ensuring the continuity, since the existing component can be replaced with minor changes in the implementation. We employ open source software Luigi for managing the distributed workflow and MongoDB for the operational and metadata databases, as examples.

The partner organizations create SIPs according to commonly agreed national specifications. A SIP is technically a directory, where the NDL METS document and a digital signature file is located in the root level, and digital objects are either in the same level or in subdirectories. The partner organization transmits the SIPs via SFTP to a buffer of the digital preservation service. The workflow manager can then find the transferred data from the buffer and start processing it. The ingest workflow has several microservices for validating the data according to the NDL specifications. Each of these microservices generate a validation report, which are finally combined as a final ingest PREMIS report.

Each AIP is packed in a Bagit format, which contains the accepted SIP and the final ingest PREMIS report. This construction is then wrapped in a compressed tar format. AIPs are finally moved to a storage buffer, which stores them to one disk media type and two different tape media types. To avoid possible technology dependency issues and/or technology lock-in, the three copies are stored with three different storage technologies produced by different manufacturers. In the first phase system, the storage media formats are: 1) Nearline SAS 4TB disks with RAID60 technology; 2) magnetic tape, Oracle T10000D drive with T10000 T2 media type; and 3) magnetic tape, IBM TS1140 drive with 3592 JC media type. We also have a dark archive service as an ultimate backup. This results that three active copies and two dark archive copies of all AIPs are stored.

Multiple servers are used for controlling the front-end, databases, disk storage, and tape libraries. The actual disk storage capacity is currently 728TB, which is distributed between five different big data oriented servers providing also processor capacity for operations in software layer, such as for ingest, dissemination, and preservation actions. Scaling out to more servers extends both disk and processor capacity. The tape libraries, in our case Oracle StorageTek SL8500 and Spectra Logic T-Finity, scale up to thousands of tapes, which makes a capacity of several petabytes possible. The tape library interfaces are open source components, and the file system employed for the tapes is a common open source standard LTFS, and therefore, the stored data can be read from the tapes even if the manufacturer discontinues supporting their own implementation. The system is connected with a redundant 10 Gbps connection to a fiber backbone network of 100 Gbps speed.

Development in hardware advances rapidly. We estimate that life cycle of the disk storage, tape drives and magnetic media types are five years, and life cycle of the tape libraries are ten years. Therefore, the hardware architecture is planned in a cost efficient way, so that we can afford to accept the fact that disks and tapes eventually get outdated, and it is business as usual to renew those. The architecture is built so that media failures do not have any effect on the preserved data nor the services, and replacing new disks is a low cost operation. Periodic bit preservation actions are also automatically performed for the data, and replacing the corrupted copies is a fully automated process.

The digital preservation service provides management tools of preserved data for the partner organization and for the system administrator. These are, for example, tools for following the ingest process or preservation actions, tools for preservation planning, and tools for updating or removing the data. The partner organizations can manage and follow only their own data. In the near future, our digital preservation service will also perform file format migrations and other logical preservation tasks to keep the data accessible and usable. This will be done in a close collaboration with the partner organizations by conforming to their preservation plans.

Our solution is dynamically scalable with practically any amount of data. The key in the scale-out technology is to distribute the storage and processor capacity between servers. Distributed file system technologies are used for this purpose, and we have chosen an open source file system GlusterFS, which takes care of the actual disk storage and a separate processing buffer for actions, such as ingest, storing, and dissemination. The connections from the partner organizations to the system are formed to a randomly selected front-end server with using a dynamic DNS bound to keepalived offered virtual IP addresses for a front-end servers available. This distributes the traffic between the digital preservation service and the partner organizations evenly between the servers.

Secure HTTPS protocol is used for all data management and access. An authenticated user can search granted data, create and download DIPs, get reports about preservation events, and get statistical information. Here, a REST interface is designed especially for automated access, and a web application using the REST interface is developed for the manual use. In the future, various additional management and access services will be developed.

Our system integrates about 40 existing (3rd party) open source components and about a dozen in-house components together. The integration work between software components, databases and services are produced in-house using the Python programming language. Selecting the 3rd party components includes an evaluation process before the actual deployment. At first, those software candidates are selected, which fulfill the needed functional requirements. The evaluation is usually done with a small but comprehensive test task, which will be implemented with the different candidates separately. These accepted candidates' maturity is then evaluated with using a method based on the QSOS (Qualification and Selection of Open-source Software) version 2 evaluation method.

We presented our national digital preservation solution concentrating on its software implementation and the hardware architecture. At the moment, the preservation services are utilized by national memory organizations preserving cultural heritage, but the services are under development for research data.

[1] H. Helin, K. Koivunen, J. Lehtonen, and K. Lehtonen. Towards Preserving Cultural Heritage of Finland. In Proceedings of the Cultural Heritage on line – Trusted Digital Repositories & Trusted Professionals, Florence, Italy, December 10–14, 2012.

[2] http://www.kdk.fi/en

Preserving Electronic Syllabi at California State University Long Beach

Chloé Pascual
CSU Long Beach
1250 Bellflower Blvd, MS 1901
Long Beach, CA 90840
1 (562) 985-1895
Chloe.Pascual@csulb.edu

ABSTRACT

Since its founding in 1977, University Archives and Special Collections at California State University Long Beach has collected syllabi from the classes offered during the academic year. Tens of thousands of syllabi exist in paper format, and thousands more exist on CD in various electronic formats. This poster describes efforts to transform syllabi collection into an all-digital process, while providing for the future preservation of historical syllabi.

General Terms

Institutional opportunities and challenges; Preservation strategies and workflows; Training and education.

Keywords

Syllabi, Access, Collection.

1. INTRODUCTION

Syllabi are an important part of building courses at any university, and they are the most accessed item from the University Archives at California State University, Long Beach. The preservation of syllabi for use by future students, alumni, and faculty touches on a number of different issues in digital preservation. Staffing issues, software issues, standard issues, and intellectual property issues all come to play in the quest to digitize and preserve the history of courses taught since the University's founding in 1949. This poster will demonstrate the state of our efforts to achieve this over the coming years.

The CSULB University Archives are unusual among CSU campuses for collecting course syllabi at all. Most university archives in the system do not collect them, and people looking for syllabi are referred to academic departments. However, at CSULB, the collection of syllabi goes back many years. Syllabi for some departments are found going back decades, and in 2004, an Academic Senate policy was passed requiring departments to turn in copies of all syllabi to the University Library by the census date of every term. In practice, collection does not reach 100%, but the University Archives still houses tens of thousands of syllabi from the roughly 18,000 course sections taught every academic year. Increasingly, these syllabi are electronic. According to Academic Senate Policy, the Library is able to specify the format of syllabi collection. In practice, University Archives and Special Collections has collected all syllabi, in any format they are given, including paper. The Archives have only asked for a different format if an electronic file was found to be unreadable. Moving forward, however, University Archives would like to streamline the collection process so that all syllabi are collected in a standard electronic format.

2. PROCESS

2.1 Evaluate

For the last year, we have been assessing our current situation. We have identified many pain points in the collection process, such as the confusion that results when syllabi are turned in multiple times during a term for a single department, or when files are labeled in unconventional ways. We have also identified vulnerabilities, such as electronic syllabi stored on CD with no copies. These are vulnerable to damage and degradation. So far, one set of syllabi has been found on a 5¼-inch floppy, and it is not yet known if there is any recoverable data still on that disc. The next step is to assess the approximate number of paper and electronic syllabi, and in what formats. We then need to evaluate the repository systems available, and possible collection methods, such as using the University's standard course management software, or perhaps using a standalone product. Once the evaluation process is done, we can create new policies and procedures for our system.

2.2 Train

Archival staff and student assistants will need to be trained on new software and procedures, and there will of course be a learning curve. Staff and faculty in other departments will also need to be trained to submit their syllabi in a new way, and for some, in a new format.

2.3 Implement

In the implementation of this project, the first step will be collecting new syllabi according to new procedures and storing them in the chosen repository. The next step will be to prepare and load electronic syllabi that are currently stored on CD in the new system. The final stage of implementation will be to digitize historical paper syllabi, so that they too are accessible electronically. This will be the most labor intensive step, but by the time it is taken, most of the difficult decisions will have been made, so it should be possible to get it done using student assistants.

3. CHALLENGES AND OPPORTUNITIES

3.1 Institutional

Special Collections and University Archives has gone through several changes in the past year. The first is that the staff member who had been managing the area retired, and a new staff member and faculty member were hired in her place. This has been a challenge, as all staffing changes are, but it also presents the department with the opportunity to update processes and procedures. Library Administration has shown a strong commitment to investing in the department, which is an important factor moving forward. University policy is also on our side with this project, as the Academic Senate passed a policy in 2004 that requires departments to turn all syllabi into the Library by the census date of each term, in the format specified by the Library. The main challenges on an institutional level are the cooperation of academic departments in changing procedures, and the limits of funds and staff time, which are always in short supply, even in a very supportive environment such as ours.

3.2 Technical

Syllabi are the most used item in the University Archives, and as such, it is important that we retain access to all of the historical syllabi collected. For this reason, the choice of format for syllabi is important. The format must be something that will continue to be supported into the future, and that will not render us dependent on a specific vendor. Additionally, the repository system chosen must be able to produce appropriate metadata so that we can search and sort the syllabi in our system as necessary.

3.3 Legal

One potential issue with providing digital access to syllabi is that many faculty members feel that syllabi are their intellectual property, and they are hesitant to allow syllabi to be available on the open web. As the CSU develops IP policies to address this issue, this may become clearer. Until that time, however, digital copies of syllabi are unlikely to be available on the open web, and are likely to be held in a closed repository for staff use only.

4. ACKNOWLEDGMENTS

The author wishes to express sincere gratitude to archives assistant Heather Steele and the student assistants of CSULB Special Collections and University Archives for their work collecting and organizing syllabi received from departments, as well as University faculty and department staff for their cooperation in syllabi collection. Library Administration has played a big role in supporting our efforts to streamline syllabi collection, and this support is appreciated as well. Finally, a very big thanks go to the author's predecessor, Kristie French, for managing an extraordinarily large collection of documents during her tenure.

Congregating Socio- Economic Datasets for Scholastic Research: A Case Study in IIMB Library

K Rama Patnaik
Librarian,
Indian Institute of Management Bangalore
Bannerghatta Road
Benguluru-560076
Tel: 91-80-26993016
Mob: 91-9740240038
Email: rama.patnaik@iimb.ernet.in

ABSTRACT

Digital curation initiatives with an intention to preserve the intellectual content of an institute gained momentum in the early 1990s with open source technologies facilitating such efforts. Indian Institute of Management Library (hitherto referred as IIMB library in this paper) heralded onto a new path of making primary datasets about socio-economic datasets, including the massive census reports in the digital domain. In the year 2014 all National Sample Surveys from the Government of India and Census reports from 1881 to 1941 were digitized for internal circulation purposes only. Though these efforts were a small step towards digital curation, it raised expectations from the user community on the computational potential and data mining abilities of these datasets. But to accomplish it, the challenges of digital perpetuity, technological obsolescence, dissemination expanded to the public, copyright issues have to be overcome.

General Terms

Institutional opportunities

Keywords

Social science data sets social surveys. India

1. INTRODUCTION

Indian Institute of Management Bangalore (IIMB http://www.iimb.ernet.in) is a leading business school located in the southern state of Karnataka in India. IIMB offers a myriad of courses spanning student population in the age group of 20 to 50 years from postgraduate courses & doctoral courses in Management to executive management courses. Its user base primarily comprises the faculty, research scholars, and students, also to substantial number of walk-in users. IIMB Library is predominantly a management resource library with more than 80% of the annual budget spent on digital resources (http://www.iimb.ernet.in/library).

1.2 Objectives

i. **Improve accessibility: T**o make the information available on the Internet. To ensure longevity of data, by digitizing content that are stored and organized in high-density servers, with searchable indexing terms for easy retrieval.

ii. **Preservation:** Preservation of original data for a longer period by the deployment of meticulous preservation techniques to protect data from deteriorating**.**

iii. **Enhance search capability:** Implement a web enabled integrated digital library through which the content can be managed, catalogued and searched.

iv. **Centre for Social Science Data Online:** Create a centre responsible for online dissemination of census data and other social science datasets, for easy and wider access.

2. DIGITAL CURATION

2.1 Digitization and Data Acquisition

IIMB Library had acquired census reports in microfiche format in 1988 and access to these reports were enabled by associated accessories such as microfiche reader and printer. The library committee amidst growing demand from the faculty of Economics and Finance area recommended for the digitization of Census reports with data mining and computational abilities. As a first step towards digital curation, these reports were converted into digital format (PDF Images) in collaboration with International Institute of Population Sciences, Mumbai

The other important datasets that were digitized were National Sample Survey Organization reports published by Ministry of Statistics, Government of India. The **National Sample Survey Organisation** (**NSSO**), now known as **National Sample Survey Office** is an organization under the Ministry of Statistics of the Government of India. It is the largest organization in India that conducts regular socio-economic surveys. Digital reports are available from 38th round onwards. However, IIMB has a print collection of earlier rounds, which were also digitized in PDF image format.

2.2 Preservation

The first stage was to preserve and make them accessible in PDF image format; we started exploring for a dedicated technology that would facilitate not only preservation but also disseminate the content. The reports were delivered in CD format; a mirroring technology was chosen to copy the files to the hard disk. Adhering to Dublin Core Schema, Metadata was granulated to include a chapter level description and make it discoverable. It also allows loading of pre-designed thesaurus unique to each type of collection, thereby allowing assignment of authoritative descriptors only. As of now, all the content is searchable, apart from browsing facility. These files will be converted into a durable format identified by the technical team for long-term preservation.

2.3 Strategies

Currently, we are evaluating from the sample PDF image files that can be converted into a data mining capable format for the tables and textual information in the reports, along with the costs incurred in the computational ability and preservation strategy.

A proposal will be submitted to the sub-committee constituted for this purpose to enable the technical team to evaluate the strategies available in the order of preference along with the cost of preservation.

a. Microfiche: Strategy one: Preserve in Microfiche either at local site or off site with a third-party vendor and digitize again as and when the current formats turn obsolete.

b. Format Migration: Reduce the risk of obsolescence by storing in multiple storage locations and then data is migrated to a new media when it is appropriate. Explore Technical registry services and digital archives projects initiated by Universities Archives to keep abreast of the formats and, software and hardware requirements 'rendering platform' by extracting the technical metadata and their durability. There are number of projects such as PRONOM of UK data archives, Jhove of Harvard University, NLNZ by National Library of New Zealand, COPTR in open planet foundations, PANIC of University of Queensland Center

c. Convert all PDF images in PDF/A-2, Use of ISO 32000-1 (PDF 1.7)

d. Tables from both census reports and NSSO data sets will be in XML from which content can be extracted in multiple formats which includes spreadsheet format. This format is more durable and adaptable to changes especially XML-based mark-up formats, with included or accessible DTD/schema, XSD/XSL presentation stylesheet(s), and explicitly stated character encoding

e. Participate and collaborate in Global archives alliances such as SafeArchive, Data–PASS and Private LOCKSS.

3. COMPUTATIONAL AND DATA MINING NEEDS:

Digitization of census reports exacerbated the demand for making these reports in a format that are downloadable and amenable for maneuvering of the numerical data. In the second phase, only the statistical data in tabular format will be made available using technology that survives obsolescence. Sample files are already being tested and it is possible to convert 90 percent of the content into XML format. We gave a few samples to evaluating the conversion process and deliverables as per our requirement.

Stage one: PDF images are converted into by PDF searchable format a by an OCR

Stage two: An application is used to convert the PDF searchable into XML.

Stage three. From this XML format, a proprietary application was run to get tables in exportable spreadsheet format. Other formats that can be generated from this are epub and HTML

4. CONCLUSIONS:

IIMB's digital curation efforts will move forward to enable the tables of census data and other social science datasets from government sources and publicly funded research projects.

An outline of the above proposal was presented to the Library Committee in the first week of September, 2015 The Committee suggested for a holistic plan to cover all digital assets of the Institute for long term preservation. This includes the Digital Institutional repositories, Electronic Journal subscriptions, Primary research datasets, MOOCs and other video lessons contributed by the faculty. A two-year time line and budgetary resources are sanctioned for the current year to initiate the digitization process for census and NSSO datasets and explore viability of strategies proposed.

ISEC (www.isec.ac.in) has digitized NSSO datasets and propose to use Dataverse Network. IIPS (www.iipsindia.org) digitized in PDF image format and offer the census content as a browsable datasets without computational and data mining capabilities. However, complete set of digitized data is currently available with IIPS and Registrar General and Census Office, Government of India from 1881 to 1991 in PDF image format. But reports of 1991, 2001 and 2011 are completely available in digital format with extractable features.

Census workstation is being set up at IIMB by the Office of Registrar General and Census office that provides access to complete published tables from 1991, 2001 and 2011.

There is a lot of potential to build theme-based collection leading to subject repositories, but our priority right now is to strengthen the socio-economic datasets. Once the data is enabled with computational and data mining abilities, we would like to make this accessible to all research scholars who are interested in this area, and extend this facility to other socio-economic data primary datasets. Preservation efforts will be simultaneously revised and implemented after the committee evaluates the strategies on efficacy and cost of maintenance. A sub-committee constituted for the purpose will evaluate after submission of the proposal for creating infrastructure for digital asset management

Assessing the Scale of Challenges for Preserving Research Data

Umar Qasim
University of Alberta
2-10N Cameron Library
Edmonton, AB
+1 (780) 492-9861
umar.qasim@ualberta.ca

Chuck Humphrey
University of Alberta
2-10T Cameron Library
Edmonton, AB
+1 (780) 492-9216
chuck.humphrey@ualberta.ca

John Huck
University of Alberta
5-25E Cameron Library
Edmonton, AB
+1 (780) 248-1337
john.huck@ualberta.ca

Leanne Trimble
OCUL Scholars Portal
130 St George St, 7th Floor
Toronto, Ontario
+1 (416) 978-7217
leanne.trimble@utoronto.ca

Alex Garnett
Simon Fraser University
Burnaby, B.C.
+1 (778) 228-5110
garnett@sfu.ca

Dugan O'Neil
Compute Canada
Burnaby, B.C.
+1 (778) 782-5623
dugan.oneil@computecanada.ca

Sean Cavanaugh
University of Saskatchewan
Saskatoon, SK
+1 (306) 966-2674
sean.cavanaugh@usask.ca

Jason Knabl
Compute Canada
36 York Mills Road, Suite/Unité 505
Toronto, ON
+1 (613) 986-0350
jason.knabl@computecanada.ca

Jason Hlady
University of Saskatchewan
Saskatoon, SK
+1 (306) 966-2075
jason.hlady@usask.ca

Rachana Ananthakrishnan
Globus Computation Institute
University of Chicago
ranantha@uchicago.edu

Kyle Chard
Globus Computation Institute
University of Chicago
chard@uchicago.edu

Jim Pruyne
Globus Computation Institute
University of Chicago
pruyne@uchicago.edu

ABSTRACT

This poster reports on the outcomes of and lessons learned from a pilot project to test core components of a national research data management infrastructure service. A software stack consisting of Archivematica and Globus Publishing was used to pass datasets from an established domain repository through an archival processing pipeline and establish discovery and access layers from the output.

General Terms

Infrastructure opportunities and challenges; Preservation strategies and workflows; Innovative practice.

Keywords

Research data, Preservation, Access, Archivematica, Globus Publishing.

1. INTRODUCTION

The number of data repositories providing access to research data is growing at a rapid rate around the globe. Developments in data access, however, have outpaced advances in the digital preservation of research data, even though long-term access is dependent on properly archived content. An important reason for this has been the wide variety of research data, its volume, and the speed at which it is produced. Finding technologies to disseminate such data tends to be easier than establishing sound ways of producing archival copies of complex datasets. Key to addressing this challenge is building software that scales to the processing demands of diverse research data collections.

As an initial investigation into this challenge in preserving research data, Research Data Canada[1] (RDC) established a Federated Data Management Pilot Project[2] to build core components of a national research data management infrastructure service. The design of the pilot project involved taking datasets from an established domain repository, passing them through an archival processing pipeline, and then establishing discovery and access layers from the archival output.

[1] http://www.rdc-drc.ca/

[2] http://www.rdc-drc.ca/activities/federated-pilot/

2. PILOT CONFIGURATION

The Canadian Polar Data Network[3] (CPDN) provided its diverse collection of research data from the Canadian International Polar Year (IPY) for use in this pilot, as well as the time and expertise of staff from CPDN partner members. A software stack consisting of Archivematica[4] for archival processing and Globus Publishing[5] for the discovery and access platform was hosted by Compute Canada[6] (CC), which also contributed personnel. The pilot project's objective was to evaluate this specific configuration to understand better the requirements for a national preservation, discovery, and access platform.

Archivematica processed each dataset selected for this pilot as a Submission Information Package (SIP) to generate Archival Information Packages (AIPs) and Dissemination Information Packages (DIPs). All DIPs were moved to the Globus Publication platform for access and discovery.

The implementation challenge with Globus Publishing was to find a flexible batch process to ingest metadata and data files from an existing collection rather than from individual research projects. This required entering metadata in batch rather than inputting metadata manually and ingesting data in bulk instead of submitting data through individual projects. Transformation of existing metadata to conform to Globus Publishing's metadata model was another key step. Aspects of this project built upon the experiences of an earlier project at Simon Fraser University by extending the deposit functionality of Globus Publishing.

3. CONCLUSIONS

This pilot provides important insights into the requirements for implementing a production service based on the functions of this test. First, it demonstrated that automated processes could generate archival digital objects for research datasets and that these objects could be deposited with an access platform (Globus Publishing in this instance) and archived in preservation storage. Second, it demonstrated that, once ingested into a discovery and access platform, datasets were discoverable and retrievable under appropriate controlled access conditions. Third, it identified a need for upfront preparation of metadata by a metadata expert and for the intervention of a data curator to start and monitor the processing cycle. Fourth, it identified several improvements that will be necessary to assemble a small-scale production system based on this pilot's basic design.

All of the suggested improvements are incremental in nature and achievable through a next-step development process. In the pilot a separate workflow was used to transmit metadata to Globus Publishing; this step needs to be better integrated into the Archivematica pipeline. Developments in computational processing that enhance scalability are needed for pushing large digital objects through the pipeline. There is a clear need for Archivematica to better manage the processing of dataset-level metadata for discovery applications outside of Archivematica. Finally, the use of Archivematica's Format Policy Registry needs to be incorporated into the design to support normalization processing of the diverse file formats encountered in research datasets.

Overall, the pilot helped us to understand better the steps required to prepare research data for access and preservation and to anticipate what a successful national preservation, discovery, and access platform for research data might look like.

[3] http://polardatanetwork.ca/

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

[5] https://www.globus.org/

[6] https://www.computecanada.ca/

Preserving Qualitative Data: A Data Model to Prepare Computer Assisted Qualitative Data Analysis Software Data for Long-term Preservation

Umar Qasim
University of Alberta
Alberta, Canada
umar.qasim@ualberta.ca

Kendall Roark
Purdue University
Indiana, USA
roark.kendall@gmail.com

ABSTRACT

A rapid change in technology has a great impact on the long-term access to digital content. This makes preservation of a digital content a challenging task due to the content's inherent dependency on a specific hardware/software platform. Changes in the technology without backward compatibility can lead to a content that can't be viewed and qualitative data is no exception. Normalization is a commonly used strategy to keep content prepared for long-term preservation. However, tools are not always available to perform normalization on complex file formats such as qualitative data analysis software files. In this work, we are proposing a data model to normalize computer assisted qualitative data analysis software files to support long term access.

INTRODUCTION

Technological obsolescence is a known phenomenon and a number of strategies have been proposed to reduce the impact of software and hardware obsolescence including normalization, emulation and migration. File format normalization is one of the preservation strategies that is being widely discussed and used in the digital preservation community. In this strategy, digital objects of a specific type are converted into a single selected format, which is thought to have a higher chance of being accessible in the future. This strategy has been used successfully with simpler file formats like text, pdf, images etc. mainly due to the availability of software libraries for normalizing these types of files. One major limitation of this strategy is that there are a large number of file formats in use and not every file type has supporting libraries available for conversion purposes. An alternate way is to do this conversion using the original application by exporting or saving the desired content into an industry standard format. Unfortunately, this process is dependent on commercial vendors to provide such a support, which is not always provided. Data driven applications such as Computer Assisted Qualitative Data Analysis Software (CAQDAS) are one example of applications which store data in complex file formats and currently no libraries are available to do the normalization process. Some of these applications are proprietary, further complicating the situation because these vendors do not always provide support for converting files into a standard file format.

PROPOSED DATA MODEL

Under these circumstances, having a deeper understanding of data models in these complex data files helps in identifying essential pieces of information needed for future access. In this poster, authors are proposing a data model approach for CAQDAS applications, which can help to extract important pieces of information whereas any gaps are covered with necessary documentation. The proposed data model is intended to provide an approach to extract and preserve all the information, which is part of a CAQDAS application file, in a way that this information can be later assembled and viewed in any other current or future CAQDAS application. Currently, some of the major CAQDAS applications lack support for interoperability amongst various CAQDAS platforms. The proposed data model provides an alternate approach to make these CAQDAS applications interoperable.

To get a deeper understanding of the whole process, Roark (2015 forthcoming) conducted one on one interviews with researchers, and Qasim and Roark (2015) conducted both a pilot and a formal workshop on documenting and preserving CAQDAS projects at the University of Alberta. During the pilot and the formal workshop, the authors demonstrated how to take a CAQDAS project apart and capture all the important study documentation embedded in the project file. Participant feedback was solicited to improve the transformation process. In addition, current preservation strategies such as normalization, migration and emulation and the contexts in which each might be used were discussed. Preservation strategies for both proprietary and non-propriety software were discussed. Furthermore, current best practices and workflows for quality assurance and documentation (metadata, provenance, codebooks, scripts) were reviewed and as well as how to operationalize ethical and contractual commitments around data access and ownership into a data management plan and preservation practices.

CONCLUSION

In this poster, we are sharing the findings of our work on preserving qualitative research data and analysis documentation. We have proposed a data model driven approach for CAQDAS file preservation and provide guidance on how to extract data model from both proprietary and nonproprietary file formats.

Protecting the Long-Term Viability of Digital Composite Objects through Format Migration

Elizabeth Roke
MARBL, Emory University
540 Asbury Circle
Atlanta, GA 30322-1006
+1 (404) 727-2345
elizabeth.roke@emory.edu

Dorothy Waugh
MARBL, Emory University
540 Asbury Circle
Atlanta, GA 30322-1006
+1 (404) 727-2471
dorothy.waugh@emory.edu

ABSTRACT

This poster documents work recently undertaken at Emory University's Manuscript, Archives, and Rare Book Library (MARBL) to review policy on disk image file formats used to capture and store digital content in our Fedora repository. Survey of the field and current best practices revealed waning support for the formats previously used and prompted collaborative efforts between Digital Archives staff and software engineers to migrate existing disk images to formats now deemed more suitable for long-term digital preservation.

General Terms

Preservation strategies and workflows

Keywords

Digital preservation; disk imaging; file format migration; PREMIS; digital archives; digital repositories.

1. THE PRESERVATION OF DISK IMAGES AT MARBL

The *Trusted Repositories Audit & Certification: Criteria and Checklist* requires that digital repositories monitor changes in technology that might impact preservation planning and maintain agile policy that can respond effectively to such changes [1]. This ongoing cycle of review and response is critical to the long-term preservation of digital objects and has been a key consideration in the development of policy at MARBL since functionality for the ingest of forensic and logical disk images was added in 2014 to our Fedora repository.

MARBL's collections include increasing numbers of digital media. A survey of the environment and best practices, conducted not long after the establishment of MARBL's Digital Archives unit, resulted in the decision to capture forensic disk images of this media using the open source Advanced Forensic Format (AFF), while logical disk images were captured using AccessData's AD1 file format. At that time, AFF offered a good solution for the capture of forensic disk images: unlike raw disk images, AFF files package disk image metadata with the image file. AFF's method of segmenting the disk image also made image compression possible [2]. That AFF is open source added to its appeal as a format for long-term preservation as it meant that we did not have to depend on limited proprietary formats, the viability of which often fluctuate in response to commercial markets. However, the development of Libewf by Joachim Metz, a library of tools supporting access to the proprietary Expert Witness Compression Formats, have decreased the need for an open source alternative like AFF. As a result, the creator of AFF, Simson Garfinkel, has stopped active development and no longer recommends AFF as a format for digital preservation [3]. In response to this shift in best practice, Digital Archives at MARBL recognized a need to update policy and workflow, which also presented a good opportunity to address the acquisition of logical disk images. Use of the AD1 file format to capture logical disk images had allowed us to generate and record fixity information as part of the imaging process. However, we were well aware that AD1's proprietary format left our data vulnerable, and we were keen to find an alternative better suited to our goals for long-term preservation.

Following conversation with colleagues across the field, we made the decision moving forward to acquire raw disk images or, where circumstances prevented complete forensic imaging, tar files. While this shift in policy did mean that we lost the benefits of the AFF format, we felt that data stored as raw disk image files was less vulnerable to obsolescence, as raw image formats are supported across platforms and their limited complexity results in a format that is, theoretically at least, more easily maintained over the long-term. Similarly, the ubiquity of tar, in addition to built-in functionality that preserves file metadata, presents a preferred alternative to the AD1 file format. While these changes are reflected in our current workflows, continued monitoring of the environment and best practices remains a key part of our policy. In response, we expect that our workflows will continue to develop and hope that they will continue to improve.

One of the challenges that resulted from this shift in workflow was how to migrate AFF and AD1 files already captured and stored in our Fedora repository.

2. THE MIGRATION PROCESS

This migration was Emory's first attempt to do file format migration in our Fedora digital preservation repository. It required us not only to develop methods for migrating the files but also to reconfigure the repository to accept new mimetypes for the newly ingested files.

The most straightforward objects to migrate were our AFF files. Software engineers performed the majority of the migration work. AFF, an open source format, has a library and toolkit that Emory used to automate the migration. Software developers obtained the AFF files from our digital repository, extracted the raw image from the AFF image, and uploaded the migrated disk image back into the repository. The fact that the AFF image itself includes the raw MD5 and SHA-1 checksums as part of the metadata enabled us to validate the migrated image upon conversion. Checksums were validated at each stage in this process to ensure the integrity of the digital object. A file containing these checksums has been stored with the object in the digital repository as a supplemental file. Once the migrated file was successfully ingested into the repository, the original AFF was deleted.

Migration of the AD1 images, a proprietary format, was a manual and much more complicated process. Fortunately, MARBL had captured only a limited number of these types of files, making this approach feasible. Software engineers obtained the AD1 files from the repository. Digital archivists loaded each AD1 file into Access Data's FTK Imager as an evidence item and extracted the files from the image. We also used FTK Imager to generate a file inventory with checksums for each file. Finally, the extracted files were packaged into a tar file using the Cygwin Console's tar utility with its built-in option to preserve file metadata.

Software engineers batch ingested the migrated files into the repository through a batch version of MARBL's normal ingest procedures. The tar files and the associated file inventories were packaged together using Python BagIt, which also generated fixity information. After the bags were ingested, the repository validated the digital object checksums and stored the file-level checksums as a supplemental file attached to the object. Once the migrated AD1 files were successfully ingested and validated, system administrators deleted the original AD1 images.

3. PRESERVATION METADATA AND SUPPORTING DOCUMENTATION

Since Emory began ingesting disk images into our preservation repository, we have relied upon the PREMIS metadata standard to encode the provenance of the original physical object. Technical metadata documenting the original environment (hardware and software) as well as forensic information about the imaging process are all recorded in PREMIS metadata. We also record events such as fixity checks. This structure is based in part on a model developed by the BitCurator project at the University of North Carolina that maps disk image metadata into PREMIS [4]. For the disk image file migration, we added a migration event to the object's PREMIS metadata that captured the details of the migration, including the software applications we used, migration dates, and other details.

Our use of PREMIS for disk images has also enabled us to capture file metadata no longer stored in the AFF or AD1 image. AFF and AD1 files natively package metadata about the original physical object and the imaging process within the disk image file. The raw disk image file format we now are using does not contain any of this valuable metadata. Instead, we are adding this information to PREMIS metadata for the object, ensuring that we are able to retain the metadata we need to preserve and access the object in the future.

4. IMPACTS

The migration of disk images from a proprietary or unsupported format to a raw file format has made it easier for us to manage and preserve these objects and mitigates the threat of obsolescence for the near term. The migration is not without long-term consequences, however. Although our extensive use of PREMIS preserves most of the metadata encoded in AD1 or AFF images, some system information captured as part of logical disk images has been lost as a result of the migration. We don't currently use any of this data, but it is a piece of forensic information about the object that we can no longer access. The deprecation of the AFF file format also means that we can no longer compress our disk images. This is not a concern now, but may be in the future as we continue to add large objects to our digital repository.

The greatest impact from migration has been on our imaging and processing workflows for composite objects. AFF and AD1 file formats, which automatically included system information and fixity information, guaranteed that that we preserved these types of objects in ways that were forensically sound. Going forward, we will be able to store the same metadata, but the process will be more complicated and require workflows that ensure we do so. Additionally, files will no longer contain any embedded metadata, meaning that we will consciously have to track that information along with the object.

The migration to a raw file format has made the digital file itself easier to preserve. The ongoing question is how easy it will be to preserve the original object it represents.

5. REFERENCES

[1] The Center for Research Libraries. 2007. *Trustworthy Repositories Audit & Certification: Criteria and Checklist,* 31-32. http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf.

[2] Garfinkel, S., D. Malan, K. Dubec, C. Stevens, and C. Pham. Advanced forensic format: An open, extensible format for disk imaging. *Advances in Digital Forensics II,* M. Olivier and S. Shenoi, Eds. FIP International Conference on Digital Forensics (Orlando, FL, January 29-February 1, 2006). Springer, New York, NY, 17-31. http://dash.harvard.edu/bitstream/handle/1/2829932/Malan_AdvancedForensic.pdf?sequence=4.

[3] AFF format deprecated. January 15, 2014. Guymager wiki. http://sourceforge.net/p/guymager/wiki/AFF%20format%20deprecated/.

[4] Chassanoff, Alexandra, Kam Woods, and Christopher Lee. Mapping Digital Forensics Metadata to Preservation Events Using Bitcurator. SAA Research Forum (New Orleans, LA, August 13, 2013). http://files.archivists.org/pubs/proceedings/ResearchForum/2013/ChassanoffWoodsLee-ResearchForumPoster13.pdf

Providing Access to Disk Image Content: A Preliminary Approach and Workflow

Walker Sampson
CU Boulder
University Libraries, 184 UCB
Boulder, CO 80309-0184
303-492-9161
walker.sampson@colorado.edu

Alexandra Chassanoff
UNC Chapel Hill
CB #3360, 100 Manning Hall
Chapel Hill, NC 27599-3360
919-962-8366
achass@email.unc.edu

ABSTRACT

In this poster, we describe a proposed workflow that can be used by collecting institutions acquiring disk images to support the provisioning of access to the born-digital content therein.

General Terms

Preservation strategies and workflows.

Keywords

Digital forensics; born-digital media; disk images.

1. INTRODUCTION

Born-digital materials are increasingly acquired by libraries, archives, and museums (LAMs). Though institutions have long been tasked with the preservation of collected materials, along with their continual access, born-digital data from removable media presents certain challenges [1]. One promising approach gaining traction among LAMs has been the adoption and use of open-source digital forensics software environments like BitCurator, for the capture and analysis of these born-digital materials [2].

However, there is currently limited support for institutions seeking to provide access to forensically captured born-digital content and associated metadata. The BitCurator Access project, which began in October 2014, seeks to address this gap by developing software to simplify access to content on raw and forensically packaged disk images.[1] In this poster, we propose a workflow that describes the capture, analysis, and final access to disk image content for collections held at the research archives at the University of Colorado Boulder.

2. BACKGROUND

The Archives at the University of Colorado Boulder has collected a wide range of floppy disk types; these reside in boxed folders or containers throughout its stacks. The Archives receives floppy disks as part of new accessions as well. While plans are in place for the implementation of both, the Archives has no software deployed which may function as a digital repository (e.g., DSpace, Fedora, Archivematica, Islandora), or collection management software deployed (e.g., ArchivesSpace, AtoM, PastPerfect).

The BitCurator Access project is currently developing BitCurator Access Web Tools (or bca-webtools) for web-based access to disk images.[2] Provision of access to both disk images and associated metadata through bca-webtools will help institutions capture and provide an access environment that reflects original order and relevant environmental context for collection materials. Additionally, the project proposes a fourth area of investigation related to access - the development of tools to aid in redacting sensitive data from disk images and other digital collections.

3. WORKFLOW

3.1 Goals and Context

The preliminary workflow described here addresses the immediate needs of the material, such as bit-level capture and triage, while remaining flexible enough to have the outputs integrate with a future digital repository and collection management software. We hope this approach allows the described methods a wider institutional applicability.

The workflow enables researchers to access a bit-level copy of a floppy disk found in an archival collection. Access is typically regarded as the last milestone of processing work, so the workflow strives for completeness to this point.

3.2 Overview of Proposed Workflow

The proposed workflow at the University of Colorado Boulder for processing born-digital materials will begin with obtaining the physical disk. The source media will be photographed and the archivist will begin the disk image acquisition process. Creation of the disk image can be performed through a number of devices, such as a USB-attached 3.5” disk drive in the case of the many IBM PC-formatted disks, or through floppy drive controllers such as the FC5025 for 5.25” disks and the KryoFlux controller in the case of either 3.5” or 5.25” disks.

Once the image has been created, a simple mount test will be run in the BitCurator environment. Floppy disk images might not mount for a variety of reasons including bad sectors, unknown file system types, or poor reads. Disks that are not mountable will be problematic for the next step, so these images will either be documented and set aside or resolved before further processing.

The BitCurator Reporting Tool will then be run to generate analytic reports on disk image content, including reporting on file formats and deleted files. The Reporting tool produces a DFMXL output through *fiwalk,* which is broadly analogous to a top-level inventory of disk image contents, and a PREMIS description. Other programs or processes can be carried out here as well, such as virus scans, and their outputs logged.

[1] http://access.bitcurator.net/index.php?title=Main_Page

[2] https://github.com/BitCurator/bca-webtools

The use of Simson Garfinkel's *bulk extractor* program, integrated with the BitCurator Reporting Tool through the *BEViewer* graphic front-end, reports on personally identifiable information and other sensitive content [3]. Information which a donor may have indicated should remain private can be discerned.

At this stage, the full context and content of the disk image is considered captured and described. The total output — disk image, logs of the disk imaging, a photograph of the media, and associated metadata and reports from the BitCurator Reporting Tool, will be placed into a single BagIt package and uploaded to a managed storage space with redundant copies.

The ability to control access to sensitive materials found on disk images is an explicit goal of the BitCurator Access project. The bca-webtools interface will use authentication at the local level to limit access to those materials flagged as containing potential PII in the previous step.

In this workflow, the aforementioned BagIt bag will become the formal archival information package (AIP) [3] in a designated repository at a future date. While implementation details are likely to change as the software develops, the workflow will place another copy of the disk image and attendant metadata in a location accessible to bca-webtools. We note here that the attendant metadata will likely be a subset, rather than a full copy, of the metadata and inventories available in the AIP. Even in the case of a disk image with no PII marked for redaction, unallocated user data extant after delete commands or overwrites may often prevent the full index present in the DFXML, or other such reports, to be available to the end user through bca-webtools. The precise relationship between the information contained in the suite of descriptive documents in the AIP, and the metadata used by bca-webtools and the end user, is subject to development.

The disk image interface can then provide access to the broader public, serving as a dissemination information package (DIP). Researchers will be able to browse and download the contents of the disk image through the software's web interface. This access point can be pointed to or indexed from a number of finding aid types, ranging from full EAD documents and library catalog entries to more custom online inventories.

4. CONCLUSION

In this poster, we describe how one institution – the University of Colorado Boulder – proposes to integrate digital forensics tools into their processing workflow to provide web-based access to disk image content. Although this poster describes a specific implementation, we anticipate that other institutions will likely follow similar steps in their workflow processing born-digital materials.

5. ACKNOWLEDGEMENTS

The BitCurator Access project is funded through the Andrew W. Mellon Foundation. The principle investigator is Christopher Lee and the co-principle investigator and technical lead is Kam Woods. Our software developer is Sunitha Misra.

6. REFERENCES

[1] Kirschenbaum, M.G., Ovenden, R., and Redwine, G. 2010. Digital forensics and born-digital content in cultural heritage collections. Council on Library and Information Resources. Washington DC. http://www.clir.org/pubs/reports/reports/pub149/pub149.pdf

[2] Lee, C.A., Woods, K., Kirschenbaum, M. and Chassanoff, A. 2013. *From bitstreams to heritage: Putting digital forensics into practice in collecting institutions.* White paper. University of North Carolina at Chapel Hill. http://www.bitcurator.net/wp-content/uploads/2013/11/bitstreams-to-heritage.pdf

[3] Garfinkel, S.L. Digital media triage with bulk data analysis and bulk extractor. February 2013. *Computer security*, 32(C):56–72.

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

ArchivesSpace-Archivematica-DSpace Workflow Integration

Michael Shallcross
Bentley Historical Library
University of Michigan
Ann Arbor, Michigan 48109-2113 U.S.A.
1-734-936-1344
shallcro@umich.edu

Max Eckard
Bentley Historical Library
University of Michigan
Ann Arbor, Michigan 48109-2113 U.S.A.
1-734-763-7518
eckardm@umich.edu

Abstract

In recent years, ArchivesSpace and Archivematica have emerged as two of the most exciting open source platforms for working with digital archives. The former manages accessions and collections and provides a framework for entering descriptive, administrative, rights, and other metadata. The latter ingests digital content and prepares information packages for long-term preservation and access. In April 2014, the Bentley Historical Library received a $355,000 grant from the Andrew W. Mellon Foundation to partner with the University of Michigan Library on the integration of these two systems in an end-to-end workflow that will include the automated deposit of content into a DSpace repository. This poster will introduce the "ArchivesSpace-Archivematica-DSpace Workflow Integration" project and its goals, strategies, and development roadmap.

General Terms

Preservation strategies and workflows; Innovative practice.

Keywords

ArchivesSpace, Archivematica, DSpace, Digital archives, Workflow development, Appraisal, Arrangement and description.

1. INSTITUTIONAL CONTEXT

The Bentley Historical Library collects and preserves unique materials related to the University of Michigan and the state as a whole. 80 years after its founding, the library has amassed 8,000 research collections that range from the papers of governors, to the records of student and faculty, to the entire historical record of intercollegiate athletics at Michigan. These holdings include more than 20 TB of digital content, with extensive web archives, born-digital archives, and digitized collections of print, photographic, and audio-visual materials. As part of its mission, the Bentley Historical Library is committed to ensuring the preservation and accessibility of this content over the long-term by implementing professional best practices and standards in its workflows and infrastructure.

The Bentley has actively managed large collections of born-digital content since the 1997 accession of former University of Michigan President James Duderstadt's personal computer. The 2010-2011 MeMail Project (funded by a generous grant from the Andrew W. Mellon Foundation) helped the Bentley develop more robust and uniform preservation procedures as staff explored strategies to collect and preserve the email of key university administrators. This work led staff to develop AutoPro, an ingest and processing tool comprised of 30 Windows CMD.EXE scripts that guides archivists through a standardized workflow and creates a full audit trail. Since moving into production in 2012, AutoPro has been used to prepare more than 230 accessions of digital archives (approximately 1.5 TB) that are accessible in Michigan's Deep Blue DSpace repository.

2. PROJECT GOALS

While an effective tool, AutoPro's command line interface and limited error handling capability create inefficiencies and the amount of time it takes to maintain and update scripts and software on individual workstations has significant implications for sustainability. To address these issues, the Bentley Historical Library sought funding from the Andrew W. Mellon Foundation to integrate ArchivesSpace, Archivematica, and DSpace into an end-to-end digital archives workflow. The unique strengths and affordances of the different systems lend themselves naturally to specific archival functions: ArchivesSpace for accessions, description, and tracking rights and administrative metadata; Archivematica for ingest and creation of Archival Information Packages (AIPs); and DSpace for preservation storage and access. In bringing these platforms together, project staff hope to achieve three main goals:

1. Streamline a digital archives workflow from ingest through the deposit of fully processed materials into DSpace. Manual interaction and intervention with digital archives will thereby be reduced to essential procedures to create greater efficiencies and remove possibilities for human error.

2. Facilitate the creation and reuse of metadata among platforms, including archival description, administrative information, and PREMIS rights. These metadata will be recorded in ArchivesSpace, associated with information packages in Archivematica (in addition to essential technical metadata), and displayed and/or acted upon by DSpace.

3. Improve reporting functionality and provide better tools in Archivematica so archivists may (a) review and

appraise files in a more thorough manner and (b) logically arrange content with archival description from ArchivesSpace (and thereby associate digital objects with archival object records).

All project deliverables, including modifications to source code, plugins, and documentation will be contributed back to the appropriate open source code bases or otherwise be made freely accessible to the archives and digital preservation communities. The Bentley will also ensure that new features and functionality are modular so that other institutions may adopt some or all of the project features (for instance, only the appraisal and arrangement tab and integration with ArchivesSpace) and/or modify code to meet local needs.

3. DEVELOPMENT ROADMAP

Artefactual Systems Inc., the developers of Archivematica, joined the project as programmers in late 2014 (having previously served as technical consultants); since then, the company has assisted with an in-depth review of the Bentley's digital collections and workflows; an analysis of existing features and functionality in ArchivesSpace and Archivematica (with additional exploration of areas for future development and integration); and the articulation of functional requirements and development priorities. System integration formally began in April 2015 with the commencement of agile development sprints. The project will proceed through the following seven phases through its deadline in April 2016:

Phase 1: Creation of a new appraisal and arrangement dashboard tab in Archivematica. Initial development will focus on creating a new Archivematica dashboard tab and user interface to characterize and appraise files before intellectually (and logically) arranging them with ArchivesSpace. This new tab will provide faceted searching within transfers and generate reports for one or more transfers (or components thereof), with information on file format and media type distribution, duplicate files, size on disk, sensitive data, etc. The Bentley also seeks improved tools for archivists to view or render content within Archivematica to gain a better understanding of intellectual content and value, confirm the presence of sensitive data, or deaccession materials.

Phase 2: ArchivesSpace integration. Once a basic user interface is established for the appraisal and arrangement tab, project staff will focus on the nuts and bolts of ArchivesSpace integration. The most prominent feature will be an ArchivesSpace pane in the appraisal and arrangement tab that will permit archivists to match files/folders from the transfer backlog with an ArchivesSpace archival object, thereby creating a Submission Information Package (SIP) that will correspond to a digital object record in ArchivesSpace and form a single 'item' in DSpace. To accomplish this, archivists will load the appropriate resource record in the ArchivesSpace pane and then navigate through the intellectual hierarchy to an appropriate level of description, at which point content will be dragged and dropped from the backlog transfer pane onto a specific archival object. In navigating the intellectual arrangement, users may create new (or edit existing) archival object records at any point, operations that will be limited to the title, date, 'level,' and a note, with other fields to be modified directly in ArchivesSpace. Once this arrangement has been finalized, the archivist will click a "Create SIP" button to initiate Archivematica ingest procedures and use the ArchivesSpace API to create or edit an archival object records and generate associated digital object records.

Phase 3: AIP repackaging. The Bentley currently stores AIPs in its DSpace repository and plans to continue using these for both preservation and access to avoid the redundancy of creating and storing separate Dissemination Information Packages (DIPs). As part of this strategy, the library packages large, multi-file digital objects in .zip files to simplify archival management and user access. Once content is moved from the Archivematica transfer backlog to an ArchivesSpace archival object, archivists will have the opportunity to select materials (within a single SIP) to be packaged together in a .zip file, with the ability to create multiple .zip files per SIP. In cases where a SIP contains only a few files, these will be deposited without packaging. After arrangement and packaging decisions are complete and a SIP has been created, Archivematica will continue with its ingest workflow, generating the .zip file(s) at the conclusion of this process. In another departure from current workflows, the METS file, metadata, and log files produced by Archivematica will be placed in a .zip file and deposited as a bitstream alongside the data in the appropriate DSpace item. With this 'chipped dog' approach, original and preservation copies of content will be available to researchers while access to the metadata and logs is restricted to archivists and repository staff.

Phase 4: Refinement of the appraisal and arrangement dashboard tab. Based upon user testing, feedback from other institutions, and additional development work, Artefactual Systems will refine the appraisal and arrangement tab to ensure that its features and layout best meet the needs of the user community at large.

Phase 5: External tools integration. Once changes to the Archivematica dashboard and integration with ArchivesSpace have been successfully implemented, Artefactual Systems will explore the integration of external tools to permit viewing and rendering a wider variety of file formats and mime types.

Phase 6: DSpace integration. Development related to DSpace will involve system-agnostic technologies such as SWORD and ResourceSync to ensure that the ArchivesSpace-Archivematica integration could be modified to function with other repositories (such as Fedora or CONTENTdm). Major requirements include the automated deposit of content to an existing DSpace collection, the crosswalk of descriptive and administrative metadata to Dublin Core elements in DSpace, and the ability to return unique 'handles' to the ArchivesSpace digital object record so that <dao> elements will include direct links to content.

Phase 7: Bug fixing / completion. The final phase of the project will involve the resolution of any bugs and final development tasks, taking into account additional user testing and feedback.

4. MORE INFORMATION

Updates on development efforts will be posted to the Bentley Historical Library's project blog, where questions and comments are welcome (see http://archival-integration.blogspot.com/).

Invitation to join the OAIS community platform

Barbara Sierman
Koninklijke Bibliotheek
PO Box 90407
2509 LK The Hague
+31 70 314 01 09
Barbara.Sierman@kb.nl

William Kilbride
DPC
York Science Park
Heslington, York YO10 5DG
+44(0)1413304522
william.kilbride@dpconline.org

Hervé L'Hours
UK Data Archive
University of Essex
Wivenhoe Park, CO4 3SQ
+44(0)1206 873 162
herve@essex.ac.uk

Paul Wheatley
DPC
York Science Park
Heslington, York YO10 5DG
+44(0)1904567891
paul.wheatley@dpconline.org

ABSTRACT

In this poster, we describe an initiative to build a community resource around the OAIS standard.

General Terms

Infrastructure opportunities and challenges

Keywords

OAIS

1. INTRODUCTION

The OAIS standard (Open Archival Information Standard) published by both the Consultative Committee for Space Data Systems (CCSDS) and as ISO14721 (last updated in 2012) has been highly influential in the development of digital preservation. As a reference model it provides a common basis for aligning disparate practice in diverse institutional settings. A range of standards have emerged around and related to OAIS including PREMIS (for preservation metadata), ISO16363 (for certification) and PAIMAS (for exchange between Producers and Archives).

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 to include cultural heritage, research data centers, commerce, industry and government.

The digital preservation community has a responsibility to keep the standard alive and relevant. The upcoming ISO review of the OAIS standard in 2017 offers a chance for a cooperative, transparent review process. It also creates an opportunity for further community building around OAIS and related initiatives.

At the 2014 4C/DPC Conference[1] in London a few people decided it was time to start thinking about the OAIS review in 2017 and the importance this standard has for the digital preservation and curation community. This group has a vision of developing an information platform around these common vocabularies, concepts, functions, and standards to develop a common view on the state of digital curation and preservation and provide the basis for a contribution to the OAIS review.

The Digital Preservation Coalition in the UK offered to host this OAIS resource which will be created, supported and maintained by the cross-domain, international digital preservation and curation community.. Every preservationist will be able to contribute, whether they are working in libraries, archives, research organizations, data centers or banking, medicine, and space agencies.

To support this platform the following will be initiated:

1. An OAIS Wiki environment
2. Exploring official review channels
3. Active interaction

2. *An OAIS Wiki environment*[2]

Feedback on a range of topics related to the use and interpretation of the OAIS standard and the related standards will provide a valuable reference for new as well as experienced preservationist. Various interpretations, related to different domains will offer insight in the actual implementation of the OAIS standard, insight that currently is not available in one place.

Some specialists in digital preservation and curation were asked to pick a topic of their interest and write a guest post about it, thus helping to start discussions.

The community will doubtless deliver a wide range of alternate perspectives and some conflicting views. The wiki will facilitate discussion and debate, and act as a record of the issues and opinions beyond those included in the formal review submission.

But this information can also be input for an editorial committee (consisting of the most active participants) to formulate recommendations which will result in a formal submission to the

[1] http://digitalpreservation.nl/seeds/the-gold-standard/

[2] http://wiki.dpconline.org/index.php?title=OAIS_Community

2017 review. In this way, the community will be able to understand all the wide ranging input to the review, maximising transparency and enabling ongoing dialogue.

3. Exploring official mechanisms

Official mechanisms for the review of ISO standards are well established via National Standard Bodies. These will be explored and described and used to give input for the upcoming review. This will support a better preparation for the review of the preservation community

4. Active Interaction

Ensuring inclusion for this large, diverse community means collaborative virtual meetings are necessary but we all recognize the value of face-to-face meetings and will seek to enable them.

The outcome from this activity is not simply a wiki nor is it a set of recommendations. The community that gathers around the OAIS standard is diffuse and fragmented.

By providing a shared open platform for the community that gathers around the OAIS we aim to ensure on-going dialogue about our standards and their implementation in the future.

In this sense the 2017 review is a milestone on the way to an engaged and empowered community rather than a destination.

Should Web Archives Be Used For Research Data Preservation?

Todd Suomela
University of Alberta
Digital Initiatives - University Libraries
Edmonton, Canada T6G 2J8
Tel # 1-780.248.1952
todd.suomela@ualberta.ca

ABSTRACT

This poster describes some of the challenges for managing web archive collections when they intersect with research data preservation. A web archive collection at the University of Alberta inadvertently harvested large data files from another institution which resulted in overloading the subscription budget for Archive-IT. A discussion followed about the appropriate policy approaches for web archive programs when they encounter research data on the web. The poster presents some of the evaluation criteria used to make decisions about including or excluding research data from web archives. Existing web archive tools are ill-prepared to deal with research data. Furthermore, responsibility for preserving research data and web documents is difficult to determine. Finally, the role of third-parties outside of the original institutions where research data is created is still unclear. Future activity in this area should address these challenges.

General Terms

Infrastructure opportunities and challenges; Preservation strategies and workflows

Keywords

web archives, research data, preservation policies

1. INTRODUCTION

In 2011 the University of Alberta library created a Circumpolar web archive collection using the subscription service Archive-IT. The new collection supplemented an existing non-digital collection which had been in place for decades, expanding into the digital realm seemed a natural extension of already existing services. Since then the web archive collection has faced a number of challenges demonstrating how digital collections present new problems and opportunities for libraries and archives.

One of the major challenges is dealing with data files published by other institutions on the web. In the summer of 2014 a one-time crawl consumed almost one-third of the annual data budget for the Archive-IT subscription service. The root cause of this was the downloading of hundreds of zip files which contained digital geographic images in the form of TIFF files. This presented a challenge for the library as a whole because the data budget for Archive-IT subsumes 13 active collections across multiple disciplines and library sub-units. For a single collection to consume over a third of the data budget is unsustainable. The key question raised by this incident is what the role of web archives is in research data management. Should web archives be considered part of the research data preservation process? How should research data be managed across varying institutions? Who should be responsible for preserving research data?

2. DECISION MATRIX FOR WEB ARCHIVES AND RESEARCH DATA

This poster presents a decision matrix for evaluating the relationship between a web archive and research data. Which questions should be asked in order to make decisions about the inclusion or exclusion of research data within a web archive repository? Is a web archive repository the right tool for backing up research data?

The following questions are part of the decision matrix for research data in web archives.

- Where is the data being stored?
  - Is it available on the open web?
  - Is it available from a reliable institution? e.g. a university, government on non-profit
  - Is there a preservation plan in place for the data?
- Is there any immediate threat for the data to be lost or be no longer available? (This partly depends on the type of institution hosting the data.)
- Who is responsible for maintaining the data over time?
  - Are there any disciplinary repositories which may be preserving the data?
  - Is there a hierarchy for data responsibility?
- Who are the future and current audiences for sustaining a copy of the data in a web archive?

  - How is access currently being managed through the open-web? Are there provisions or licenses which may affect access to the data through other sources, such as a web archive?
  - What provisions or tools are there available for metadata management within the web archive toolset?

- Are there resources available for storing this data in the web archive?
  - Can the data be stored in the web archive given existing data budgets, metadata descriptive services, and staff budgets?

- What is the value timeline for this research data?
  - Can a web archive preserve enough information to make this data useful for researchers in the future?
  - Will this data be useful for researchers at your institution in 10 or 20 years?

3. CONCLUSION

The evaluation process for collecting research data through web archives is still underway at the University of Alberta libraries. Early conclusions from the process are presented below:

First, existing web archive software is ill prepared to preserve research data. Most web archiving tools are designed to capture the entirety of files published by a web domain or seed with minimal filtering. The controls for metadata creation are rudimentary, the presentation layer focuses on the fidelity of web browser presentation instead of information retrieval, and the collection management interfaces allow for very limited filtering or description of file types. Given the variety of file types associated with research data web archiving should not be the first choice for research data preservation. Specific data repositories, such as Dataverse, at local institutional or disciplinary levels, make much more sense.

Second, determining the responsibility for preserving the web, whatever the form of content which is posted, is currently quite difficult. Existing data repositories are covered by a diverse range of preservation policies. Marcial & Hemminger [1] conducted an online survey of 100 scientific data repositories and identified preservation policies for 62% of the sample, but the particular policy content was idiosyncratic. Preservation information or policies are even harder to identify for individual web sites or domain names. The only feasible way to preserve large volumes of the web and have a layered approach to preservation is to develop an automated description of preservation policies similar to the robots.txt files used to limit web crawling.

Third, the role of third-parties in preserving research data from institutions with which they do not currently share any explicit agreements is very challenging. Libraries may be willing to preserve research data from other institutions but these projects are often expensive, especially in the amount of personnel time needed to coordinate the activities between multiple institutions. Most web archiving programs do not have the resources to pursue such in-depth agreements for preserving research data. This means that decisions about collecting research data within web archives will continue to be a singular challenge.

4. ACKNOWLEDGEMENTS

Thanks to Chuck Humphrey and Umar Qasim for discussing this proposal.

5. REFERENCES

[1] L. H. Marcial and B. M. Hemminger. Scientific data repositories on the web: An initial survey. *Journal of the American Society for Information Science and Technology*, 61(10):2029–2048, October 2010.

Making the Pieces Fit: Integrating Preservation into a Digital Material Ecosystem

Jennifer L. Thoegersen
University of Nebraska-Lincoln
P.O. Box 884100
Lincoln, NE USA 68521
+1 402 472-4558
jthoegersen2@unl.edu

ABSTRACT

In 2014, the University of Nebraska-Lincoln (UNL) Libraries implemented the Rosetta preservation software to take a more proactive approach to the preservation of digital materials for which UNL Libraries are stewards. A significant part of this process was determining how to integrate this new software into the Libraries' current workflows and ecosystem for digital materials. This included considering the varied origins of digital materials; determining the purpose of collections and whether preservation was necessary and, if so, a priority; considering deposit strategies; understanding the rights related to the materials, including whether or not they should be accessible and to whom; and implementing policies that outline in which systems digital materials of certain types should reside.

This poster will illustrate UNL Libraries' progress toward implementing Rosetta in its digital material ecosystem. A diagram presenting the relevant technical 'pieces' and accompanying explanatory text will demonstrate how preservation—more specifically, Rosetta—fits into the puzzle of digital material storage, access, and management. In addition, the poster will reflect on the challenges encountered attempting to fit all of the pieces together.

General Terms

Preservation strategies and workflows

Keywords

Digital preservation, libraries, implementation

1. INTRODUCTION

The University of Nebraska-Lincoln (UNL) Libraries host a wide array of digital materials which come from varied sources and serve diverse purposes. Materials include instructional materials, research data, digitized and born-digital archival and special collections, web-based projects, and multimedia collections. For many of these materials, the Libraries have an interest in—and often mandate to—act as stewards and ensure the content's long-term preservation. While the Libraries' kept multiple backups of all digital materials, there was no active digital preservation initiative in place to ensure the integrity of and continued access to digital materials far into the future. With this in mind, the Libraries' Data Curation Committee began investigating digital preservation options and selected the Rosetta preservation system by Ex Libris.

2. DIGITAL ASSET ECOSYSTEM

2.1 Content Streams

The Libraries both generate and curate digital content. The University Archives & Special Collections department collects ephemera related to the University, as well as prominent Nebraskans. Retiring faculty occasionally deposit their research materials with the university on separation from the University, as do University Chancellors and system presidents. The mission of the Archives is "to select, preserve, arrange, describe, provide reference assistance for, and promote the use of rare and unique research materials. The Department maintains these research materials because they are best managed separately from the general collections due to their subject area, rare or unique qualities, source, physical condition or form."[1] In addition to undertaking projects to digitize their unique collections, Archives collects born-digital items from other campus entities.

The Center for Digital Research in the Humanities, a joint program between the University Libraries and the College of Arts & Sciences, generates digital, often web-based research projects with varying rates of continued development. These projects will often have source material, e.g. high resolution images, as well.

UNL Libraries manages one of the largest institutional repositories in the country, DigitalCommons@University of Nebraska-Lincoln[2], which includes Master's theses and doctoral dissertations, peer-reviewed journals/series, and 'works', both published and unpublished, generated by researchers affiliated with UNL.

UNL Libraries Image & Multimedia Collection[3], powered by CONTENTdm, includes approximately 100 collections totally over a quarter of a million items spanning over a dozen disciplines. The purpose of collections vary from instructional to archival to historical. Most items are images, with a smaller percentage of text-based and audio formats. Files are generally of web quality, though, depending on the purpose and origin of the collection, items may have inaccessible high quality equivalents, as well. Item deposit is decentralized, occurring at branch libraries and a variety of units across campus.

[1] http://libraries.unl.edu/archives-special-collections-mission-collection-scope

[2] http://digitalcommons.unl.edu/

[3] http://contentdm.unl.edu/

UNL Libraries maintains a data repository, where researchers can deposit datasets to be managed by the Libraries and to provide public access to their data.

2.2 Fitting in Preservation

A major challenge in implementing a preservation system was determining how it would interact and complement or replace existing systems. Emphasis centered on what content should go into Rosetta, in what order it should be ingested, and how.

The question of what materials to include in the digital preservation system proved more challenging than originally anticipated. Many of the digital materials maintained by the Libraries are not items the Libraries necessarily have much interest in preserving. For example, a large number of the collections in CONTENTdm include low resolution images for which UNL does not hold the copyright. Instead, we are able to provide access to the campus community for instructional purposes. Working through the approximately one hundred collections in CONTENTdm and assessing the rights, the content quality, and whether the collection's stewardship and preservation falls within our mission is essential prior to ingest items into Rosetta.

The order of content to be ingested was largely determined by the following factors:

- **Readiness:** The readiness of both the collection and the relevant unit was a major factor in determining first collections to ingest. Collections needed to be well-organized and have consistent metadata to help ensure quality mapping. Units needed to be able to allocate adequate staff time to be trained on using Rosetta, assist in developing a workflow, ensure the quality of the initial ingest, and continue deposits with minimal intervention
- **Volume & homogeneity:** In order to quickly ingest a large amount of content at once, large collections that were similar enough to use the same workflow were selected. For example, Archives & Special Collections has many collections of digitized images with consistent metadata. An initial focus was to develop a workflow to allow for regular, continued deposits with little variance.
- **Necessity:** All content selected for preservation needs to be ingested in Rosetta; however, some collections may have a more immediate need, e.g. improved accessibility.

How content would be ingested was the final and most complex hurdle to getting content into Rosetta. While the variety of content complicated the issue, the variety of ingest options was most challenging. Exploring the various setup options was time consuming especially when attempting to identify the implications for these choices, e.g. how much staff time in each department will be required both upfront and ongoing, what training will be necessary, how expansion will occur as more units begin using the system, what method makes the most sense for legacy and future data, what technical issues may hamper the workflow.

The variety of options has allowed for tailored solutions for units:

- CONTENTdm export converted to Rosetta METS packages for automatic ingest
- Ongoing single-item manual deposits
- Manual bulk deposits
- ZIP deposit with CSV metadata

3. DISCUSSION

Integrating a new system into an existing setup can be a challenge, especially when attempting to create new workflows to transfer data seamlessly. Pieces will not fit together perfectly, and bridges must be built to span the gaps. This step, from creation to Rosetta ingest, is the first of many along the road of long term preservation.

4. ACKNOWLEDGMENTS

My thanks to DeeAnn Allison, Michael Litwa, Jason Bougger, Mary Ellen Ducey, Joshua Caster, and the various individuals and units at UNL Libraries that assisted in Rosetta implementation.

Targeting Audiences among the Masses: A Data Curation MOOC for Researchers and Information Professionals

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

Thu-Mai Christian
Odum Institute for Research in Social Science
228E Davis Library, CB# 3355
Chapel Hill, NC 27599-3355
(919) 962-6293
thumai@email.unc.edu

Rachel Goatley
University of North Carolina School of Information & Library Science
3201 Davis Library, CB# 3355
Chapel Hill, NC 27599-3355
(919) 843-0998
rgoatley@live.unc.edu

ABSTRACT

In this poster, we illustrate the work of the IMLS-funded Curating Research Assets and Data Using Lifecycle Education (CRADLE) project in developing a data curation massive open online course (MOOC) targeted to two distinct audiences: researchers who are becoming increasingly burdened with data management policies, and information professionals tasked to support these researchers. The poster describes data curation concepts selected for its applicability to both audiences as well as how content and delivery of educational materials are varied to enable students to achieve learning objectives.

General Terms

Training and education

Keywords

Massive open online courses, MOOCs, Data curation, Data management, Training, Education

1. POSTER SUMMARY

Of the many activities involved in creating resources for data curation education, translating data curation concepts into terms and ideas that are relevant and understandable to different target audiences has been a central concern for the IMLS-funded *Curating Research Assets and Data Using Lifecycle Education* (CRADLE) project. The CRADLE project is currently in its second year of producing high-quality educational materials focused on data management best practices for both researchers and the information professionals who support these researchers. The product central to the CRADLE project is a massive open online course (MOOC), which provides an educational content delivery platform with the potential to reach a global audience of individuals—albeit with vast differences in prior data curation knowledge, involvement in data curation activities, and perspectives on data curation.

In developing the MOOC, the CRADLE project team considered these conspicuous differences as they identified topics essential to the understanding and application of data curation concepts—whether the learner is a researcher or an information professional. These topics are:

- Understanding research data
- Data management planning
- Working with data
- Sharing data
- Archiving data

Each of these topics are being packaged as individual MOOC modules with lecture videos, interviews with data curation experts and researchers actively participating in data management activities, multimedia illustrations of concepts, quizzes, practical exercises, discussion board prompts, and other supplementary learning materials.

At the same time, the CRADLE project team understood that each of its audiences warrants different teaching approaches, with variations in explanations of data curation concepts, examples, and exercises to more effectively deliver concepts in a manner that resonate with individual students. Because the development of a MOOC requires a significant amount of research, planning, labor, and resources, the CRADLE project needed to formulate a single MOOC that would enable both researchers and information professionals to achieve learning objectives.

This formula combines delivery of relevant content through the Coursera on-demand course platform that allows students to select modules most relevant to their specific interests and learning needs; diverse perspectives from information professionals and researchers from various disciplinary domains reflected in lecture discussions and supplementary materials; and adaptable exercises that allow students to apply data curation concepts to practical situations they are more likely to encounter in the real world.

This poster outlines the essential data curation topics to be covered in MOOC modules, as well as how the delivery of MOOC content considers the distinct learning objectives of both the researcher and the information professional audience. It will also present next steps for the CRADLE project as it works to achieve its broader goal of establishing networks of data curation education and practice.

2. ACKNOWLEDGEMENTS

The CRADLE project was made possible in part by the Institute of Museum and Library Services, under award #RE-06-13-0052-13. Members of the CRADLE project team are Dr. Helen Tibbo (Principal Investigator), Thu-Mai Christian (Co-principal Investigator), Jonathan Crabtree, Rachel Goatley, Barrie Hayes, Michele Hayslett, Sophia Lafferty-Hess, Kasha Mammone, and Danianne Mizzy

Strategies for Audit-based Repository Certification: Guidelines, Resources, and Tools to Prepare, Organize, and Evaluate Criteria Evidence

Jessica Tieman
National Digital Stewardship
Residency
202-512-0000 x30992
jtieman@gpo.gov

ABSTRACT

In this poster, we present the current status, lessons learned, and best practices experienced thus far in the preparation for audit and certification of the Government Publishing Office's FDsys as a Trustworthy Digital Repository. The poster will serve as an introduction to a future, publically accessible toolkit and set of resources and case studies for use within repositories seeking an audit-based approach of evaluation.

Keywords

Standards and practice, models, preservation action and planning, risk mitigation, risk management, archival storage, national approaches, audit, certification, government

1. INTRODUCTION

The Government Publishing Office (GPO) launched the Federal Digital System (FDsys) in January 2009 as a content management system, preservation repository, and public website providing access to legislative, executive, and judicial publications including the Congressional Record, Federal Registrar, and Federal appellate, district, bankruptcy, and national court opinions. Since its inception, FDsys has received over one billion document retrievals, and stakeholders are eager for FDsys to pass an external ISO 16363 audit.

The Audit and Certification of Trustworthy Digital Repositories (TDR) superseded the Trustworthy Digital Repositories Audit and Certification (TRAC) and became ISO standard 16363 in 2012. Since this time, only five digital repositories have been established as "TRAC-Compliant." In addition, the TDR checklist has become a basis for certification in accordance to ISO 16919 as of November 4, 2014.

During this same time, the Library of Congress and the Institute of Museum and Library Services have made the 2015-2016 National Digital Stewardship Residency (NDSR) possible. NDSR serves to build a designated community of professionals who will advance the nation's capabilities in managing, preserving and making accessible the record of human achievement held in digital form [1]. The Government Publishing Office has been awarded one Resident from the NDSR program to assist in the preparation for audit. With the added support of this resident, GPO's first steps in the preparation for an audit includes conducting an in-depth gap analysis of artifacts mentioned in the TDR checklist versus existing GPO artifacts. Following this step, the GPO Resident will reach out to at least four certified TRAC-compliant repositories that follow the OAIS reference model but have not pursued certification as either a TRAC-compliant or Trustworthy Digital Repository and gather feedback about the audit and certification process. The internal audit conducted by the Resident and the GPO FDsys Trustworthy Digital Repository project team is planned to be completed in May 2016. At this time, this poster will serve to summarize the current progress of the internal audit activities and knowledge gained thus far to share with the digital stewardship community.

2. OBJECTIVES

The first step for any repository seeking an audit-based approach to certification is to accomplish an internal gap analysis to address inconsistencies or required documentation related to policy and practice. Before pursuing an external audit, an organization must be strategic and specific in accurately assessing if current policies and practices are truly sustainable, consistent, and adequate for such an evaluation. While completing an internal gap analysis through the summer of 2015, the Government Publishing Office's Trustworthy Digital Repository project team will conduct a review of resources, tools, and services that repositories might consider while conducting an internal audit.

This presentation serves to support the overarching mission and values behind the NDSR program. NDSR projects provide an experiential learning opportunity for post-graduate digital preservation professionals. The Library of Congress Digital Preservation and Outreach and Education initiative has previously surveyed a variety of digital preservation organizations and institutions across the country and found that over half of respondents expressed a need for additional staff dedicated to their digital preservation efforts [2]. To meet the educational objectives of the NDSR experience, this poster will serve as an informational presentation of the NDSR Resident's experience at the Government Publishing Office. The poster hopes to act as a guide for other emerging professionals and suggest strategies during preparation for standards-based certification.

3. References

[1] Library of Congress. 2015. National Digital Stewardship Residency. Webpage. http://www.digitalpreservation.gov/ndsr/

[2] Library of Congress. 2014. *Digital Preservation Outreach and Education Program: 2014 Training Needs Assessment Survey Executive Summary*. Technical Report. http://www.digitalpreservation.gov//education/2014_Survey_Executive_Summary-Final.pdf

Using the Virtual-Private Cloud Model to Serve and Preserve Historical Collections: A Case Study (Based on Islandora)

Gail Truman
Truman Technologies, LLC
4096 Piedmont Avenue, Ste. 217
Oakland, CA 94611
+1510-502-6497
gail@trumantechnologies.com

Jaime Henderson
California Historical Society
678 Mission Street
San Francisco, CA 94105
+1415-357-1848 X214
jhenderson@calhist.org

ABSTRACT

This poster session describes the selection criteria and process used for evaluating three repository software offerings and cloud platforms, with pros and cons. It describes implementation of workflows, representations of PREMIS metadata for objects in the repository, documenting fixity checks performed on datastreams, mapping of "rights" elements in DC datastreams to PREMIS "rightsExtension" elements, and more.

General Terms

Infrastructure opportunities and challenges

Keywords

storage cloud, Islandora, digital repository, SOAR®, preservation[1]

1. INTRODUCTION

The California Historical Society sought to implement a digital asset management and repository system to help preserve and showcase two terabytes of digitized materials. Faced with aging on-premise servers and storage, the society decided to remove the financial and resource burden of technology migration and local IT staffing and move from capital expense to an operation expense model– one based on a virtual-private, secure cloud.

2. PHILOSOPHY APPLIED

When evaluating and recommending approaches to the long-term protection of digital assets, we apply the following "big rules" or philosophies

- *Keep it simple*: Digital repository systems should be easy to implement, understand, and support.
- *Don't overbuild*: If you try to anticipate every "what if" scenario, you will a) overspend, b) be late to deploy, c) probably not need everything that was included.
- *Lots of copies keep stuff safe*: Ideally store 3 copies of all data in 3 different geographies, stored 3 different ways.
- *Have an exit strategy*: Standards-based open software and SLAs provide for vendor liquidation or end of services.

3. REQUIREMENTS FROM CALIFORNIA HISTORICAL SOCIETY

The historical society had a high-level list of requirements coming into the projects that are shown below. After a detailed evaluation of the three options, additional and more granular, requirements were identified. These requirements were given a weight for their importance and then given a score.

- Bulk ingest
- Fixity check
- Metadata standards
- Cloud based
- Stable URLs
- Bit preservation
- Exit strategy
- Rights and restrictions
- Public interface
- Exposed to Google
- Offers staff training
- Includes support
- Open source
- Cost
- Others are using it
- User friendly viewers for books, manuscripts, photos, and other formats
- Streaming
- OCR
- Online exhibitions
- Restrictions/embargos
- Analytics/stats
- Multi-lingual

4. OUR SOLUTION

Truman Technologies, LLC (TTL) utilizes Islandora software and the open-source SOAR® (Scalable Online Archive & Repository)* stack to recommend the best possible digital repository solution for organizations weighing their digital options. For the California Historical Society, TTL recommended the Islandora DAM repository software as offered and serviced by Discovery Garden Inc. (DGI), hosted by the secure private cloud (Infrastructure as a Service) vendor KomodoCloud. TTL also recommended that another copy of all digital assets (historical collections) be hosted at DuraCloud.

5. BENEFITS OF ISLANDORA

Aside from meeting California Historical Society's functional requirements, there were several preservation features that were drove the decision to move forward with Islandora:

- Preservation File Formats and Normalization: a single workflow can ingest a preservation/archival file and convert it into a customized preferred format as well as an access format.
- Versioning[2]: a record of how digital content has changed allows clients to preserve the data as well as the look and feel of a digital object, since its original dissemination mechanism and all subsequent changes are stored and linked to the original content.
- Interoperability, Reusability, and Bitstream/ Object Replication[3]: all data from a digital object are stored in a format that can be exported to future systems or shared between applications.
- Fixity, File Format Identification, and Data Integrity: Islandora FITS, Checksum, and Checksum Checker modules add functionality to the Islandora solution pack by adding technical metadata extraction, enabling checksum generation, and adding a PREMIS "fixity check" entry to an object's audit log[4].
- Preservation Metadata: the PREMIS module[5] produces XML and HTML representations of PREMIS metadata for objects in a repository; its current features include documenting all fixity checks performed on datastreams, including "agent" entries for a given institution and for the Fedora Commons software, and mapping contents of each object's "rights" elements in DC datastreams to equivalent PREMIS "rightsExtension" elements.
- Community Interest Groups: Islandora has an active Preservation Interest Group[6] that develops and recommends common approaches to preservation within the Islandora suite.

[1] SOAR (Scalable Online Archive & Repository) is a trademark of Truman Technologies, LLC and is registered in the US Patent and Trademark Office.

[2] Davis, D. (2011, August 13). Versioning- Fedora 3.4 Documentation.

[3] Access and view the BagIt module, documentation, and forum at https://github.com/islandora/islandora_bagit

[4] The checksum and checksum checker modules are community developed, and not yet supported by DGI, but future versions of Islandora will include these features. Access the checksum and checksum checker modules at https://github.com/Islandora/islandora_checksum and https://github.com/mjordan/islandora_checksum_checker

[5] Access the PREMIS module, documentation, and forum at https://github.com/islandora/islandora_premis

[6] Read the Preservation Interest Group's manifesto at https://docs.google.com/document/d/1o6GeNbFQsR5L_4BBe1UtwbORcRnW0ii0SXoGdEDfR98/edit#

Preserving Informal Astronomy: Arceli, the PressForward Plugin, and the Archiving of Scientific Communications

Stephanie Westcott
Roy Rosenzweig Center for History
and New Media
George Mason University
4400 University Drive, Fairfax, VA
westcott.chnm@gmail.com

Kelle Cruz
Arceli, Chair of the Board
ScienceBetter Consulting
4 West 101st Street, #29
New York, New York
kellecruz@gmail.com

Eric Olson
Roy Rosenzweig Center for History
and New Media
George Mason University
4400 University Drive, Fairfax, VA
eric@pressforward.org

ABSTRACT

This poster will profile and demonstrate a new collaboration focused on community curation, preservation of digital communications, and the archiving of science information on the open web. The PressForward Project, a research initiative concerned with the discoverability of digital gray literature, including blog posts, white papers, data visualizations, and podcasts, and Arceli, a collaborative effort within communities of astronomers whose mission is to preserve informal astronomy communications, are in the process of developing a method to make it possible to curate, archive, index, and cite digital alt-publications.

The PressForward Project, funded by the Alfred P. Sloan Foundation and based at the Roy Rosenzweig Center for History and New Media, was founded in 2011 in response to driving questions about the ability for researchers in any field to keep up with the growing body of literature—some peer-reviewed, most not—found on the open web. As the quantity of gray literature regularly published online continues to grow, so does the task of locating and evaluating the work most relevant to a particular area of scholarly investigation. How does any one scholar or community do it? PressForward seeks both technological and methodological solutions to this problem.

PressForward's technological innovation is the PressForward plugin, a free, open-source WordPress plugin that provides a smoothly integrated editorial process for the aggregation, review, discussion, and republication of external web content within the WordPress dashboard. PressForward aggregates content via RSS feeds, functions as a seamless feed reader, and allows users and groups to mark and discuss individual items before modifying or reproducing them for republication. Communities use this technology to discover and share the most relevant, highest quality work found online, an approach taken by PressForward's prototype publication Digital Humanities Now. This poster will document the use and workflow of the plugin, and presenters will have laptops on hand to demonstrate the plugin and allow audience members to test the plugin in action.

Arceli was created in recognition that the Internet allowsastronomers to publish useful material outside of traditional journals, yet there is no effective mechanism for these communications to be archived, indexed and cited. Arceli acts as a facilitator for archiving communications on behalf of authors and providing a structure that NASA/SAO Astrophysics Data Service (ADS) can index. Arceli is envisioned as an archive of alt-publications relevant to the professional astronomy community, including tutorials and how-tos, editorials and commentary, and research blogging.

As a PressForward partner, Arceli envisions using the PressForward plugin to facilitate the curation of materials that will then be submitted to Zenodo, an online repository, where it can be assigned a DOI and be indexed by the ADS. By using these tools, Arceli will preserve the informal communications many astronomers rely upon and make them discoverable and citable. This poster will preview the method that Arceli is currently developing, visualizing the process that a blog post or tutorial will go through from publication to aggregation and curation using PressForward to storage and DOI assignment in Zenodo and indexing in ADS.

The collaboration between PressForward and Arceli represents an innovative approach to the long-term preservation of informal scientific communications, one that we hope can be adapted by other communities with similar archiving needs. The poster will bring together an illustration of the plugin's features and Arceli's methods, offering an example of a creative workflow from publication to preservation, helping audience members better understand the possibilities of this approach for their own projects.

General Terms

Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice.

Keywords

digital preservation, gray literature, curation, indexing, citation

Research on Accessibility of Digital Documentation on Physical Media across Different Versions of MS Windows

Shunsuke Yamamoto
Digital Library Division, Kansai-Kan,
The National Diet Library
8-1-3 Seikadai, Seika-cho,
Soraku-gun, Kyoto, Japan 619-0287
(+81) 774981482
s-yamamo@ndl.go.jp

ABSTRACT

This poster presentation describes the results of a research project conducted by the National Diet Library (NDL), which investigated the accessibility of digital documentation stored on physical media across different versions of operating systems. This project was conducted from 2012 to 2013 as a part of a larger research project to investigate the practicality of long-term preservation and use of digital library materials stored by the NDL on physical media.

General Terms

Preservation strategies and workflows

Keywords

Long-term accessibility, Media Collection, Digital Preservation

1. BACKGROUND

The National Diet Library (NDL) is the sole national deposit library in Japan and is responsible for developing and preserving a comprehensive collection of material published in Japan as part of the cultural heritage of both the present and future generations.

Since the NDL Law was amended in 2000, the NDL has been collecting digital material stored on physical media such as floppy disks, MOs, CD-ROMs, DVD-ROMs, USB flash drives and Blu-ray Discs under the legal deposit system. The NDL classifies these digital materials into three categories: audio material (e.g. audio CDs), video material (e.g. films on DVD), and digital documentation that is neither audio nor video.

There is a great variety of digital documentation that cannot be classified as either audio or video. For example, applications that retrieve corporate information from databases stored on CD-ROMs, residential and other special-purpose maps, archival databases of past issues of newspapers and magazines, supplements to monographs, software programmes, and conference proceedings. The NDL also collects video games. This digital documentation comes on a wide variety of physical media, including optical discs such as CD-ROMs, CD-Rs, CD-RWs, DVD-ROMs, and Blu-ray discs, as well as magnetic storage media, such as floppy disks and USB flash memories. The majority of digital documentation held by the NDL, however, is stored on optical discs, and therefore we specifically focused on optical discs in this research.

As of FY2014, the NDL had acquired some 121,000 optical discs and continues to acquire roughly 8,000 new items each year.

Digital documentation is only accessible via service terminals (Windows PCs) located in the reading rooms at the Tokyo Main Library or other NDL facilities. There are also some stand-alone terminals used for viewing certain materials that must be isolated from any network for security reasons. The digital documentation of most interest to patrons appears to be archival databases of newspapers, specialized maps, databases of securities reports or other corporate information, and conference proceedings.

2. OBJECTIVE

During 2011, the NDL completely replaced its integrated library system, including the service terminals in reading rooms. At the same time, the operating systems for service terminal were changed from Windows XP to Windows 7. This upgrade caused some trouble in terms of the accessibility to digital documentation. As a result of this change, a significant number of digital documents were no longer readily available for use due to incompatibility with the new operating system.

To better understand this issue, the NDL conducted a research project from 2012 to 2013, which examined the accessibility and usability of digital documentation by determining which versions of the operating system were needed for the digital documentation to function properly.

3. METHOD

For this research, 52 digital documents were selected as test documents. 21 of the test documents were documents that had been reported as having failed to install or execute properly on the Windows 7 service terminals. Another 31 materials were chosen at random based on publication date, versions of Windows, or other attribute. All 52 documents were stored on optical discs.

A breakdown of the 52 test document is as follows: 7 documents were designed to operate on Windows 3.1, 6 items on Windows 95, 4 items on Windows 98, 1 item on Windows ME, 8 items on Windows 2000, 13 items on Windows XP, 8 items on Windows Vista, and 5 items on Windows 7.

These documents were tested in the following environments: Windows 7 (32 bit), Windows 7 (64bit), Windows XP (SP3), Windows 2000, and Windows 95.

We attempted to open each test document in each of the environments listed above from the latest (Windows 7) to the oldest (Windows 95), and rated each document as being playable, playable with additional application attached to the material, partly playable, or unplayable. Some documents were rated as unplayable, because an additional application needed to open the document was not attached.

Each document was tested to see if: the necessary files could be installed, the document could be opened, the document could be used properly (including searches or operations), and the document could be displayed properly (text, graphic image, video, audio).

4. RESULT

The final results of this evaluation process are shown in Table 1. The figures shown represent a percentage, calculated by dividing the number of documents that passed the test by the total number of documents. There were 8 documents that were excluded from the figures for Windows 2000 and Windows 95, because they were not tested due to hardware problems.

Table 1. Percentage of documents that were compatible with each environment

	Playable	Playable with additional application	Partly playable	Unplayable
Win 7 (32)	48%	10%	8%	35%
Win 7 (64)	38%	12%	8%	42%
Win XP	65%	6%	13%	15%
Win 2000	59%	5%	11%	25%
Win 95	45%	7%	5%	43%

Major issues encountered during testing are as follows.

-Some documents were installed successfully but failed to run properly. In one case, installation completed properly, but the start menu never appeared.

-Some documents required specific playback applications other than those normally packaged with Windows computers, or specific plug-ins for a specific application.

-Some documents required a specific framework for installation and execution. In one case, Microsoft .NET Framework 1.1 was required.

-Some documents were compatible with only one version of Windows. (Primarily, Windows 7 (64 bit))

-Some documents were not installed due to missing files or information.

5. CONCLUSION

We found that Windows XP was compatible with more documents than any other version of Windows. We assume that this is related to the fact that Windows XP was intended to be compatible with both MS-DOS-based Windows 9x and Windows NT-based products.

We also found significant incompatibility between 32-bit and 64-bit versions of Windows. We found many documents that would either not install or not run properly on the 64-bit version of Windows 7.

Before conducting this research, we assumed that the newer documentation would have greater accessibility and usability. The results of this research, however, demonstrate that this is not entirely true. In fact, the older test documents, which were created on Windows 3.1, had a higher level of compatibility than newer documentation. We speculate that, since this documentation was published in the early days of physical media, it does not have the more complex programmes or strict copy protection used for documentation that was published in the later years.

Still, this study was conducted on a relatively small number of test documents, and these results could change significantly if conducted with a different set of test documents.

Apart from issues with the different versions of Windows, there were other factors that resulted in errors during installation or running of these documents. Some documents need specific applications to install or operate properly, such as Acrobat Reader, Quicktime, or Shockwave. During this research, there were several cases in which the necessary applications were not supplied with the original materials. Therefore, these materials could neither be installed nor executed.

There was also some documentation that required serial numbers, passwords or product keys for installation or activation. The problem here is that such information is frequently not readily accessible. For instance, in one case, the necessary information was stored on floppy disks, which are now obsolete. Since the service terminals are not equipped with 3.5” floppy disk drives, it was difficult to extract the necessary information.

The result of this research suggests that, in order to maintain a proper playback environment, it is essential to preserve not just the actual documents but all passwords, product codes, application software, and other items needed for installation and execution of the digital documentation.

We also investigated availability of the same content online or in alternative media, and found that little of it was readily obtainable.

6. REFERENCES

[1] *The Long-term accessibility of packaged digital publications (NDL Research Report No.6)* http://www.ndl.go.jp/en/aboutus/dlib/preservation/research_report2006.html (retrieved on 2015/6/5)

Automatic Identification and Preservation of National Parts of the Internet Outside a Country’s Top Level Domain

Eld Zierau
Department of Digital Preservation
The Royal Library of Denmark
Søren Kierkegaards Plads 1
DK-1016 København K
ph. +45 91324690
elzi@kb.dk

ABSTRACT

Preservation of our cultural heritage on the Internet is increasingly in danger of getting lost due to the challenges faced when collecting it. An increasing amount of national webpages are moving to generic Top Level Domains like .com or .org. The movement is so fast that we are at risk of losing it, since we do not get in time to identify the change before it has disappeared again. Therefore this question becomes increasingly crucial for organizations covering digital national heritage including web archives for a specific country.

This poster presents the results from a research project that evaluated two different automated approaches to recognise webpages outside a country’s Top Level Domain which are part the country’s cultural heritage. One suggested approach has been to base extraction of national material on a snapshot of the entire Internet in form of a worldwide crawl. Another suggested approach is more silo oriented, based on harvests of web pages referred to by webpages within a National Top Level Domain.

More specifically the research project aimed to identify automatic procedures for evaluating the two suggested approaches, and for identifying Danish web content on websites outside the national Top Level Domain “.dk”. The datasets used were links from a 30TB Danish 2012 bulk harvest and the 360 TB Internet Archive *wide-0005 crawl,* since these two harvests are comparable in time frame.

The poster will present

- The two methods and the difference in their results
- Indications that the two approaches find very different material
- The general method used to evaluate the nationality of web material over time

The general method mentioned here is important, since the very basis for any harvesting approach is defining a collection scope by deciding what is seen as national webpage. Automation of such definitions is far more difficult than originally anticipated. The automation here is based on a wide range of general criteria that are implemented (e.g. language recognition, national terms like ‘je suis Charlie’ or phone number patterns). An additional outcome of the project has been a generally applicable list of collection criteria, which is based on a cooperative effort between representatives within the fields of scholarship, the Danish web archive, and computer science.

General Terms

Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice.

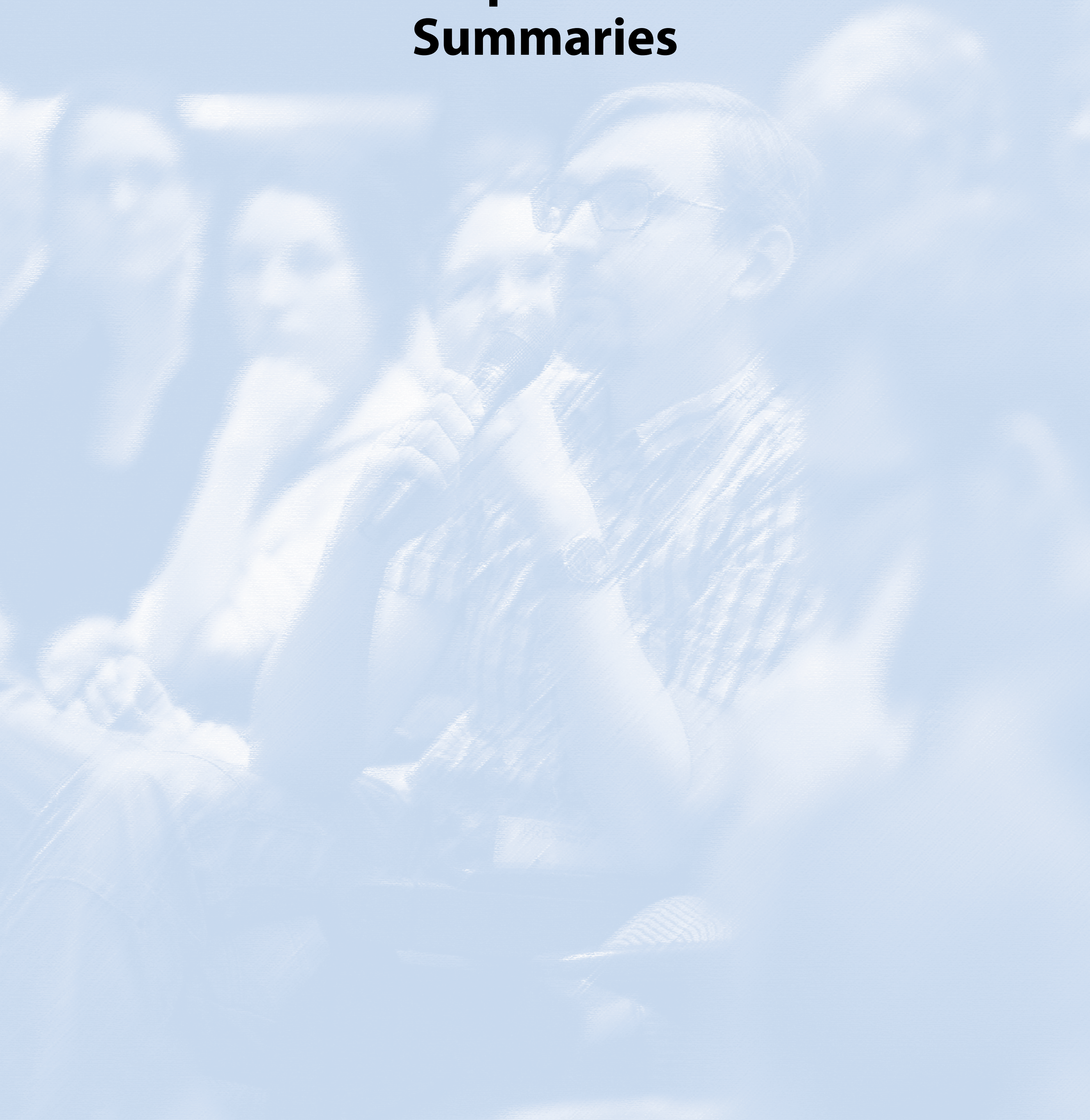
Keywords

Preservation, web archive, collection strategies

Note

It should be noted that the last mentioned criteria method part has been presented at the RESAW 2015 conference, but in a closed forum, - and the first part with the results have been presented at the IIPC GA in a presentation, but not as a poster which opens a better possibility to discuss and understand in depth, as well as exchange ideas based on the results.

Workshop and Tutorial Summaries



Data Mining Web Archives

Jefferson Bailey
Internet Archive
300 Funston
San Francisco, CA, CA 94118, USA
jefferson@archive.org

Lori Donovan
Internet Archive
300 Funston
San Francisco, CA, CA 94118, USA
lori@archive.org

ABSTRACT

Many institutions are now building rich, significant archives of web content. Though the number of web archiving programs has grown, access models for these collections have remained focused on URL-based discovery and traditional live-web-style browsing. Given the resources required to build and maintain web archives, finding new forms of access for these collection will help increase use and thus allow institutions to better advocate for the value of collecting and preserving web content.

Distant reading, text mining, digital humanities, and other data-driven forms of analysis have become increasingly popular methods of using digitized and digital collections. Web archives, being born-digital, of notable size and temporal breadth, having extensive metadata, and often created with a curated topical focus, are ideal resources for data mining and other forms of computational analysis.

This workshop will explore new methods of research use of web archives by giving attendees exposure to, and training in, the tools, methods, and types of analysis possible in working with datasets extracted from the entirety of curated web archive collections. Giving researchers datasets of specific extracted metadata elements, link graph data, named entities, and other post-processed data can help facilitate new uses and new types of visualization, inquiry, and analysis.

Workshop Objectives:

- Introduce attendees to web archives and the issues of provenance, formats, methods of collection, and the core tools and technologies involved in web archiving
- Give an overview of the types of derived datasets that can be created from web archives
- Provide sample datasets, scripts and tools, and outline research and use scenarios
- Explore methodological challenges and possibilities
- Lead attendees through a data analytic workflow that includes processing, publishing, and visualizing web archive data

Keywords

Web archiving, data mining, research, access

Testing the Proposed METS 2.0 Data Model against Use Cases and Complementary Data Models: Presentations and Community Discussion

Bertrand Caron
Dept. of Bibliographic and Digital Information
Bibliothèque nationale de France
Quai François Mauriac
75706 Paris Cedex 13, France
Bertrand.caron@bnf.fr

Andreas Nef
Docuteam GmbH
Im Langacker 16
CH-5405 Baden-Dättwil
Switzerland
a.nef@docuteam.ch

Nancy J. Hoebelheinrich
Knowledge Motifs LLC
448 East Ellsworth Court
San Mateo, CA 94401, USA
+1-650-302-4493
nhoebel@kmotifs.com

Thomas G. Habing
Library Software Development Group
155 Grainger Engineering Library
Information Center MC-274
University of Illinois at Urbana-Champaign, USA
thabing@illinois.edu

ABSTRACT

The Metadata Encoding and Transmission Standard (METS) 1.x schema has an established community of users including academic and national libraries, archives, and museums as well as support from a number of commercial and open source tool and service vendors. While the established community of METS users has adapted systems and tools to METS expressed in XML, many in the library and archive communities are moving toward the use of newer technologies such as those of the Semantic Web and linked data for the digital content that they have been collecting. As a result, the METS Editorial Board (MEB) has been contemplating a data model for a next generation METS schema that will facilitate these kinds of technologies. The initial approach to a new METS data model aligned very closely to metadata schemes in the preservation arena, namely PREMIS, but the MEB thought it essential to test the new METS 2.0 data model against existing canonical implementations of METS, and developing complementary data models. This workshop will describe current and ongoing efforts to evaluate and further develop a new METS data model. Participants are invited to participate in the discussions, and the subsequent evaluation / refinement of a METS 2.0 data model.

General Terms

Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice.

Keywords

Aggregation formats; Digital object description; Metadata Encoding and Transmission Standard, Metadata standards alignment, Digital preservation.

1. INTRODUCTION

The Metadata Encoding and Transmission Standard (METS) 1.x schema has an established community of users including academic and national libraries, archives, and museums as well as support from a number of commercial and open source tool and service vendors. While the established community of METS users has adapted systems and tools to METS expressed in XML, many in the library and archive communities are moving toward the use of newer technologies such as those of the Semantic Web and linked data for the digital content that they have been collecting. In order to accommodate the interests of this community and anticipate the needs of the established METS community, the METS Editorial Board (MEB) has been contemplating a data model for a next generation METS schema that will facilitate these kinds of technologies.

2. INITIAL APPROACH

One of the most common, and canonical uses for METS has been to facilitate the preservation of digital objects in libraries, archives and museums, often in conjunction with PREMIS. In addition, because the PREMIS standard had already been transformed into RDF, the MEB thought that an initial approach to a new METS data model should include a close alignment to other metadata schemes that were compatible and complementary, such as PREMIS and OAI-ORE. The first draft of the METS 2.0 data model, introduced in 2014, was built with these kinds of alignments in mind. Subsequent MEB and community discussion resulted in the desire to test the new METS 2.0 data model against existing canonical implementations of METS, and developing complementary data models, especially those for structuring both simple and complex digital objects.

3. APPLYING USE CASES

Use cases from two canonical implementations of METS have been developed to provide a better understanding of how the first draft of the METS 2.0 data model could be applied to both a relatively simple digital object and a more complex, 3D object. In the course of the application to the use cases, a number of issues have arisen, such as what elements and attributes (from the XML perspective) would be important to keep or adapt from the METS 1.x schema, which could or should be used from other schemas, and the implications of the possible choices. While the two initial use cases have not yet been fully developed, the MEB would value community input upon the findings to date.

4. THE BROADER CONTEXT`

While the METS Editorial Board has been developing a next generation data model, other, similar efforts have arisen by other communities in the libraries, archives, and museum communities. The PREMIS Editorial Committee continues to work on a version 3.0 of a PREMIS ontology that promises useful symmetry with METS as does, potentially, efforts to craft the Portland Common Data Model that is being developed by the Duraspace community. As collaboration seems more productive than competition in the area of digital object aggregation and description, speakers knowledgeable about complementary data models will discuss their data models and what issues have arisen that could benefit from cross-format collaboration.

5. THE PROS / CONS OF STANDARDS ALIGNMENT

Other issues have arisen as a result of the application of use cases, and the exploration of ways to adapt the proposed METS 2.0 data model. While the choice of RDF as a means of serializing a data model permits the re-use of classes and properties from another schema, there appear to be some disadvantages to this approach that give pause. For instance, it does seem important to keep in mind the overall purpose or goal of a complementary schema to more fully anticipate the implications of the re-use of classes and properties. Given the overall purpose of a complementary schema, when and for what reasons is it more advisable to create a new class within a METS domain than re-use one from another domain? Community discussion of these issues will be solicited using specific examples that have arisen from the application of the use cases to the METS 2.0 data model.

Benchmarking Forum

Kresimir Duretec
Vienna University of Technology
Karlsplatz 13,
1040 Wien,
Austria
duretec@ifs.tuwien.ac.at

Artur Kulmukhametov
Vienna University of Technology
Karlsplatz 13,
1040 Wien,
Austria
artur.kulmukhametov@tuwien.ac.at

Andreas Rauber
Vienna University of Technology
Karlsplatz 13,
1040 Wien,
Austria
rauber@ifs.tuwien.ac.at

Christoph Becker
Vienna University of Technology & University of Toronto
27 King's College Cir,
Toronto, ON M5S,
Canada
christoph.becker@utoronto.ca

ABSTRACT

This proposal is for a full day workshop to be held at the IPRES 2015 conference. The focus of the workshop will be on areas in the digital preservation field which could benefit from benchmarking. Benchmarking is a method of comparing entities against a well-defined standard (i.e. benchmark). The workshop is focused on discussing software benchmarking practices in digital preservation and how these can contribute to improving digital preservation tools

General Terms

Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice.

Keywords

Digital preservation, benchmarks, benchmarkDP, evaluation, software tools, empirical evidence, workshop.

1. INTRODUCTION

The Digital Preservation field is characterized by a variety of tools used to collect and process digital information. The quality of those tools is of great importance to the preservation community. However, quality assessment is often done in an isolated way independent of systematic and community driven initiatives. This results in a lack of solid evidence to support quality claims.

Benchmarking is a systematic method of facilitating comparisons of software artifacts according to a well-defined standard (benchmark). It has shown itself as a valuable empirical method for evaluating software tools. Various fields such as software engineering and information retrieval have reported major benefits from properly established benchmarking initiatives.

Several of the most successful, such as TREC[1] and CLEF[2], have annual meetings where they report on benchmark results, refine already existing benchmarks and define new ones. However the overall successfulness of benchmarking is dependent on the readiness of the community to accept and drive the whole process.

There are several indicators showing that the digital preservation field and the community are mature enough and ready for establishing benchmarking as a community driven method. This would generate benefits, such as provision of solid evidence around the quality of digital preservation tools to enable improvements of those tools [2].

The goal of this workshop is to bring together key stakeholders in digital preservation to discuss the needs of benchmarking in the field, and to define and prioritize initial benchmarks. The workshop is imagined as the first workshop in a series of workshops where community members would report on benchmark results, define new and refine already existing benchmarks.

It is distinct from the Capability Assessment and Improvement workshop at IPRES 2013[3] which focused on the assessment and improvement of organizational capabilities [1]. In contrast, the focus of this workshop is on benchmarking software tools.

2. WORKSHOP FOCUS & OUTCOMES

The workshop is planned as a full day event. We would like to see participants from various domains with a common interest in software quality challenges in digital preservation. We will encourage submission of short position papers (two pages maximum) defining a specific benchmarking need, articulating the

[1] http://trec.nist.gov/

[2] http://trec.nist.gov/

[3] http://benchmark-dp.org/caiwipres/

challenges behind that benchmark and the benefits expected from implementing that benchmark in the community.

The workshop will be divided into two parts. The first part will consist of:

- An introduction session, where we give an overview of the workshop together with a brief introduction into benchmarking as a method, with illustrations and experiences from other fields;
- A session where participants present their benchmark proposals and perspectives. Each presentation will be accompanied by time for discussion.

It is expected that these two sessions will fill the time until lunch. After lunch we will focus on providing a discussion forum to brainstorm and discuss possible benchmarks. This will build on the morning session and will include defining each benchmark according to a common structure with defined components. In the case of a bigger number of submissions, proposals will be grouped and worked on in breakout sessions and each group will present their outcomes.

In the final wrap-up session, we plan to discuss the workshop outcomes, reflect on the possibility and challenges of establishing benchmarking in the digital preservation field, and distill specific steps forward and a future roadmap.

All the proposals submitted by the participants will be published online before the workshop. Following the workshop, we will write a workshop report for the D-Lib Magazine.[4] Finally, each defined benchmark will be released in a full specification published online to enable wider community involvement.

3. INTENDED AND EXPECTED AUDIENCE

This workshop will be open for all participants. Certain categories of expected participants can be listed, however:

- Stakeholders whose digital preservation operations are dependent on the high quality of software tools, who know specific aspects that could benefit from benchmarking. We encourage the proposal of real world scenarios considered important to the community.
- Tool developers familiar with a specific domain (e.g., validating PDF files or component integration in preservation systems).
- Researchers with an interest in software benchmarking, testing and test data generation are expected to bring fresh ideas for building certain benchmarks and addressing the challenges defined by other participants.

4. ORGANIZERS`

Kresimir Duretec is a Project Assistant at the Department of Software Technology and Interactive Systems (IFS) at the Vienna University of Technology. He is currently pursuing his PhD at the same department. Previously he graduated with an MSc and BSc in Computer Science from the University of Zagreb in 2011 and 2009 respectively. He previously worked as a Sub-project lead of the SCAPE Planning and Watch sub-project. Currently he is working on the project BenchmarkDP, where his main focus is on benchmarking digital preservation tools and automatic test dataset generation using model driven engineering principles.

Artur Kulmukhametov is Project Assistant and PhD student at the Software and Information Engineering Group, Vienna University of Technology. He was involved in the European research project SCAPE. His current focus in the project BenchmarkDP is the systematic evaluation of software tools for digital preservation, content profiling and quality assurance of migration processes.

Andreas Rauber is Associate Professor at the Department of Software Technology and Interactive Systems (IFS) at the Vienna University of Technology (TU-Wien). He furthermore is president of AARIT, the Austrian Association for Research in IT. His research interests cover the broad scope of digital libraries and information spaces, including specifically text and music information retrieval and organization, information visualization, data analysis, neural computation and digital preservation.

Christoph Becker is an Assistant Professor at the University of Toronto where he leads the Digital Curation Institute, and a Senior Scientist at the Software and Information Engineering Group at Vienna University of Technology in Austria. He was involved in the European research projects DELOS, PLANETS, DPE, and SHAMAN. He led the sub-project Scalable Planning and Watch in the SCAPE project and he is Principal Investigator of BenchmarkDP. His research focuses on digital libraries, digital curation and digital preservation, and sustainability in software engineering and information systems design.

5. PROCESS FOR CONTRIBUTION

The workshop is an interactive event focusing on discussions and inspirational exchanges between participants on software benchmarking, related challenges and opportunities in digital preservation. To achieve this goal, we will publish a call for participation and ask participants to submit a short abstract with their experiences and thoughts on benchmarks they wish to see discussed at the workshop. The review criteria for these submissions will be based upon for the overall suitability of the topics proposed.

6. ACKNOWLEDGMENTS

The workshop is supported by the Vienna Science and Technology Fund (WWTF) through the project BenchmarkDP (ICT12-046).

7. REFERENCES

[1] C. Becker and E. Cardoso. Report on the Capability Assessment and Improvement Workshop (CAIW) at iPres 2013. D-Lib Magazine, 20(3/4), Apr. 2014.

[2] K. Duretec, A. Kulmukhametov, A. Rauber, and C. Becker. Benchmarks for digital preservation tools. In Proceedings of IPRES 2015, 2015.

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

PREMIS Implementation Fair Workshop

Evelyn McLellan
President, Artefactual Systems
201 – 301 Sixth Street
New Westminster, BC
Canada, V3L 3A7
+1 604 527 2056
evelyn@artefactual.com

Karin Bredenberg
National Archives of Sweden
P.O Box 12541
SE-102 29 Stockholm, Sweden
+46 10 476 71 23
Karin.bredenberg@riksarkivet.se

Rebecca Guenther
Library of Congress/Consultant
101 Independence Ave SE
Washington, DC 20540 USA
+1 703 298 0157
rguenther52@gmail.com

ABSTRACT

This workshop provides an overview of *the PREMIS Data Dictionary for Preservation Metadata*, a standard addressing the information you need to know to preserve digital content in a repository. It includes a brief introduction to PREMIS and the launch of version 3.0, which changes the data model. In addition there are reports from the preservation community on implementation of the standard in various systems or contexts, in particular the integration of preservation systems that support PREMIS with other digital management systems.

General Terms

infrastructure, preservation strategies and workflows, case studies and best practice, digital repositories, preservation repositories

Keywords

Preservation metadata, Preservation repository implementation, Data dictionary

1. INTRODUCTION

The PREMIS Implementation Fair Workshop is one of a series of events organized by the PREMIS Editorial Committee [1] that has been held in conjunction with previous iPRES conferences.

At iPRES 2015, the workshop will give the audience a chance to understand the PREMIS data dictionary's new version 3.0 and give implementers, and potential implementers, of the *PREMIS Data Dictionary for Preservation Metadata* an opportunity to discuss topics of common interest and find out about latest developments.

2. OUTLINE OF WORKSHOP CONTENT

2.1 Overview of the revised PREMIS Data Dictionary

The *PREMIS Data Dictionary for Preservation Metadata* [2] is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. Developed by an international team of experts, PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into a number of commercial and open source digital preservation tools and systems. This session provides an overview of the PREMIS Data Model as it has been revised in version 3.0. This revision widens the scope of the types of objects that PREMIS may be used to describe, allowing for the description of intellectual entities. It includes an enhanced ability to describe the components of environments comprised of software and hardware, thus expanding the scope to include physical objects. The session explains the model behind how environments are handled in version 3.0; software and hardware may be described as agents responsible for events and also may be part of compound objects that describe an environment which itself must be preserved.

2.2 Integrating PREMIS into the wider digital ecosystem

There are various efforts underway to integrate systems that provide different roles and functions in the management and preservation of digital objects. This session reviews how different types of digital preservation and management tools, such as ArchiveSpace, Archivematica, DSpace, BitCurator and Islandora, might work together to aggregate or exchange PREMIS metadata and provide a more comprehensive digital preservation solution.

2.3 Implementation reports

Reports from the preservation community will discuss implementation in various standards and contexts. Representatives from the Netherlands Institute for Sound and Vision and the Museum of Modern Art in New York will discuss their PREMIS implementations. Others in attendance are encouraged to present the choices they have made in implementations and issues encountered, engendering discussion from the participants. Discussion often results in proposed revisions of the PREMIS standard.

3. WORKSHOP SERIES

The PREMIS Implementation Fair at iPres 2015 will be the seventh in a series that have been held in conjunction with iPres since 2009. These events are intended to highlight PREMIS activities, discuss issues concerning implementation, and provide a forum for implementers to compare their activities, issues and solutions. Because this is a rapidly changing area, it is important to provide continuous updates.

4. INTENDED AUDIENCE

The workshop is designed for those involved in selecting, designing or planning a preservation project or repository using preservation metadata. This includes digital preservation practitioners (digital librarians and archivists, digital curators, repository managers and those with a responsibility for or an interest in preservation workflows and systems) and experts on digital preservation metadata and preservation risk assessment.

5. SHORT BIOGRAPHIES OF ORGANIZERS

Karin Bredenberg works as a technical advisor in metadata at the National Archives of Sweden. She is at the same time responsible for the work with creating national profiles of standards used in e-archiving in Sweden. She is a member of the PREMIS Editorial Committee as well as the METS Board. She further acts as a member of the Technical Sub-Committee on EAC-CPF, the Technical Sub-Committee on EAD and the Schema Development Team with the Society of American Archivists.

Evelyn McLellan is President of Artefactual Systems Inc., a Canadian company which develops open-source software for archives, libraries and museums. She is a member of the PREMIS Editorial Committee and the Conformance Sub-Committee.

Rebecca Guenther has been Chair of the PREMIS Editorial Committee on which she has served since its establishment in 2006. She worked at the Library of Congress on metadata standards in the Network Development Office for 22 years and is currently an independent consultant in New York on metadata development and training. She also continues to work periodically for the Library of Congress. She was co-chair of the original PREMIS Working Group which developed the *PREMIS Data Dictionary for Preservation Metadata.*

6. PROCESS FOR SOLICITING CONTRIBUTIONS

Contributions will be solicited from the PREMIS Implementers' Group via its discussion list (pig@loc.gov). To subscribe go to: http://listserv.loc.gov/listarch/pig.html. The PREMIS Editorial Committee will review all requests. If this workshop proposal is approved, then a call will be sent for contributions to the implementation portion, the deadline for which will be within a month.

7. REFERENCES

[1] PREMIS Maintenance Activity, http://www.loc.gov/standards/premis/

[2] PREMIS Editorial Committee. 2012. *PREMIS Data Dictionary for Preservation Metadata* (Library of Congress). http://www.loc.gov/standards/premis/v2/premis-2-2.pdf

Using Open-Source Tools to Fulfill Digital Preservation Requirements

Courtney Mumma
Artefactual Systems
+1-604-527-2056
courtney@artefactual.com

Michael Shallcross
University of Michigan
+1-734-936-1344
shallcro@umich.edu

Sam Meister
Educopia Institute
+1-404-783-2534
sam@educopia.org

Christine Di Bella
ArchivesSpace
+1-678-235-2905
christine.dibella@lyrasis.org

Bradley Westbrook
ArchivesSpace
+1-678-235-2910
Brad.westbrook@lyrasis.org

Christopher A. Lee
University of North Carolina
+1-919-962-7024
callee@ils.unc.edu

Max Eckard
University of Michigan
+1-734-763-7518
eckardm@umich.edu

ABSTRACT

This workshop offers a space to talk about open-source software for digital preservation, and the particular challenges of developing systems and integrating them into local environments and workflows. Topics will include current efforts and grant-funded initiatives to integrate different open source archival software tools; the development of workflows involving multiple open source tools for digital preservation, forensics, discovery and access; and the identification of gaps which may need filled by these or other tools.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice; Training and education.

Keywords

Open source; workflows; case studies; demonstrations

1. INTRODUCTION

The past decade has seen the development of numerous open-source tools that can support the curation of digital collections. Many professionals are looking for further guidance on how and when to adopt particular tools, in order to support their institutions specific needs, constraints and workflows.

This proposed full-day workshop in the innovative practice stream will (1) help to raise aware of the features and capabilities of these tools, (2) highlight cases in which the tools are being used (in isolation or together) in specific settings, and (3) facilitate focused discussions on how to best adopt and integrate such tools.

2. INTENDED AND EXPECTED AUDIENCE

This workshop is designed for those with direct digital preservation responsibilities; those with strategic and oversight responsibility related to digital preservation; and those engaged in tool development and testing.

3. OUTLINE OF WORKSHOP

This workshop will be composed of several sessions:

- Personal introductions and overview
- Tool demonstrations
- Case studies - discussion of workflows at specific institutions, including gaps in tools and how those are being addressed or could be addressed
- Panel discussion of lessons learned from demos and case studies (including integration endpoints, existing gaps and potential coordination options)
- Small group discussion – participants will break into smaller groups to discuss implications for their own institutions and opportunities for collaboration
- Reporting out – groups will report out about their discussions and the larger group will convey what they see as (individual and collective) next steps

4. OVERALL ORGANIZING STRATEGY

The workshop organizers will issue a call for participation. Interested parties should submit a short summary (one page maximum) of a demonstration or case study they would like to present. These contributions will serve as the basis for the tool demonstration and case study portions of the day. The workshop organizers were serve as panelists during the third portion of the day, and they will then serve as facilitators for the break-out group discussions.

5. WORKSHOP ORGANIZERS

- Christine Di Bella, Community Outreach and Support Manager, ArchivesSpace
- Max Eckard, Assistant Archivist for Digital Curation, Bentley Historical Library, University of Michigan
- Christopher A. Lee, Associate Professor, School of Information and Library Science, University of Michigan
- Sam Meister, Preservation Communities Manager, Educopia Institute
- Courtney Mumma, Consultant, US and International Community Development, Artefactual Systems
- Michael Shallcross, Lead Archivist for Curation, Bentley Historical Library, University of Michigan
- Bradley Westbrook, Program Manager, ArchivesSpace

Roles and Responsibilities for Sustaining Open Source Platforms and Tools

Trevor Owens
U.S. IMLS
1800 M St. NW
Washington, DC, 20036
tjowens@imls.gov

Carl Wilson
Open Preservation Foundation, c/o
The British Library, Boston Spa,
Wetherby, West Yorkshire, LS23 7BQ
carl@openpreservation.org

ABSTRACT

Developing, deploying and maintaining open source software is increasingly a core part of the core operations of cultural heritage organizations. From preservation infrastructure, to tools for acquiring digital and digitized content, to platforms that provide access, enhance content, and enable various modes for users to engage with and make use of content, much of the core work of libraries, archives and museums is entangled with software. As a result, cultural heritage organizations of all sizes are increasingly involved in roles as open source software creators, contributors, maintainers, and adopters. Participants in this workshop shared their respective perspectives on institutional roles in this emerging open source ecosystem. Through discussion, participants created drafts of a checklist for establishing FOSS projects, documentation of project sustainability techniques, a model for conceptualizing the role of open source community building activities throughout projects and an initial model for key institutional roles for projects at different levels of maturity.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges;

Keywords

FOSS, Sustainability, Institutional Roles,

1. INTRODUCTION

As cultural heritage institutions become increasingly involved in collaborative development, deployment and maintenance of open source software an ecosystem of researchers, non-profit organizations, cultural heritage institutions, service providers and funders have emerged to help make this work possible. The roles and responsibilities that these entities should take are often only evident in the successes of individual open source tools and platforms. Through facilitated discussion, participants in this workshop focused on formalizing the kinds of roles that these organizations can and should play in developing, deploying, sustaining, and disseminating open source software, tools, best practices, and services.

2. PARTICIPANTS

There were 35 participants in this day-long workshop. Attendees brought their experience working in a range of roles at a variety of institutions. There were participants from the Computer History Museum, the National Library of Sweden, the National Archives of Australia, the Bentley Historical Library, the State Archives of North Carolina, Artefactual Systems, Educopia Institute, and the John F. Kennedy Presidential Library. Participants also represented a cross section of roles (administrators, archivists, librarians, lawyers, software developers, and community managers) within organizations. This diversity of backgrounds, roles and perspectives provided invaluable input, leading to fruitful discussion.

3. WORKING GROUP OUTCOMES

The attendees organized themselves into four working groups. These groups began drafting guides and resources to address a range of pressing needs for improving investment and planning for FOSS digital preservation projects. The work of the groups is briefly described below.

3.1 Checklist for Establishing FOSS Projects

Where does one start when planning a successful open source project, or open sourcing an existing software project? While there is some work related to the maturity of FOSS projects [4] there is still a significant need for the guidance in this area. Recognizing the complexity in this space, one group began drafting a checklist for key issues to consider and explore when considering starting an open source project or shifting an existing software development project to an open source model. The group identified a range of individual issues organized into five categories; planning, legal and licensing, requirements and testing, user community, and developer community. When revised and completed, this checklist will be useful as a resource to both establish plans and also as a tool to evaluate plans for proposed tools.

3.2 Identifying Sustainability Techniques

Establishing approaches to address the sustainability of FOSS digital library projects remains a key issue area in the field [5]. There are various modes for generating the funds or in kind contributions necessary to make an open source software project sustainable. Through discussion of a range of individual projects and of related research this group articulated a series of techniques for sustainability and noted their strengths and weaknesses. Through this process, the group produced a set of notes highlighting key features of successful open source projects. In particular, participants noted that most mature open source projects in the digital library sector leverage core operating resources across multiple organizations. The group also noted that the most successful projects incorporate multiple streams of funding and resources helping to ensure sustainability.

3.3 FOSS Community Building Planning

The success of open source software projects is anchored in their ability to engage and develop communities [1]. Through discussion of the development of successful and vital open source communities around a range of different individual projects this group began articulating critical community building activities. These activities are tied to different stages in a project (from conceptualization, to design and development, through to implementation and adoption). A key take away from the group is the importance of establishing community development plans at every stage of a project's development. There is a clear need to complete the development of this model to clarify and share which activities are appropriate at particular stages of a project.

3.4 Organizational Roles & Project Maturity

This group examined and discussed different, successful open source software projects. They defined a set of project phase, identifying key roles for different institutional partners during the development of these projects. This suggested the following roles over three distinct phases of development.

Key roles identified for the initial development/ *start-up phase* of a product were:

- Researchers/Developers working to document needs, explore possibilities of tool creation;
- User Stakeholder groups working to develop use cases and features, as testers and as initial testers;
- a Steering Committee, made up of key individuals who can ensure institutional commitments; and
- clearly identified stakeholders working on documentation [2].

As a project reaches it's *initial roll out and moves toward maturity*, it becomes important to engage:

- professional associations (to get the word out about the project),
- a sustainable home (an organization focused on running and managing the project and providing services, managing membership models, and serving as a host for member driven governance).

When a product *reaches maturity*, it ideally will have cultivated:

- other providers (companies and or non-profits providing additional services around the product); and
- a developer community (a community of developers from multiple organizations contributing to the project.

4. CONCLUSIONS & NEXT STEPS

Each group identified some next steps and key participants that plan to carry forward the work started in the meeting. This will fully realize the development of resources and guides that can be used to improve the planning, delivery and quality of open source software for digital library and digital preservation tools and systems. In closing remarks, Paul Wheatley of the Digital Preservation Coalition, stressed how critical it is for knowledge and best practice in this area to advance. Every year significant resources are invested in software development across the sector. Without further development of the kinds of resources started by these working groups, it is difficult to ensure that those investments are making the maximum impact.

5. ACKNOWLEDGMENTS

Many thanks to all 35 of the participants who worked in this session, as well as the iPres conference for providing a venue for this work.

6. REFERENCES

[1] Bacon, J. 2009. The Art of Community. O'Reilly Media.

[2] Blankenhorn, D. 2010. Why Open Source Documentation Lags. *ZD Net.* http://www.zdnet.com/article/why-open-source-documentation-lags/

[3] Doran, M. 2007. "The Intellectual Property Disclosure Process: Releasing Open Source Software in Academia" http://code4lib.org/2007/doran

[4] Reese, T. (2012). Purposeful Development: Being Ready When Your Project Moves From 'Hobby' to Mission Critical Issue 16, 2012-02-03 http://journal.code4lib.org/articles/6393

[5] Schaefer, S. 2010. Challenges in Sustainable Open Source: A Case Study" Issue 9, 2010-03-22 http://journal.code4lib.org/articles/2493

Curating Research Assets and Data Using Lifecycle Education (CRADLE): Curation with a Focus on Preservation

Helen R. Tibbo
School of Information and Library Science
Chapel Hill, North Carolina, USA
tibbo@ils.unc.edu

Thu-Mai Christian
Odum Institute for Research in Social Science
Chapel Hill, North Carolina, USA
tlchristian@unc.edu

ABSTRACT

As major funding agencies, publishers, and research institutions continue to issue data sharing, management, and archiving policies in increasing numbers, libraries are being called upon to support researchers in their efforts to comply with these policies. To be responsive to researchers' data needs and to increase the likelihood of effective and efficient data preservation, many data librarians and archivists are seeking the knowledge, skills, and competencies necessary to confront the growing—and increasingly complex—data management and preservation needs of their institutions. With lecture, discussion, and hands-on exercises, this tutorial will explore the obligations of researchers to manage their data, identify the attributes of data that add to the complexity of data curation tasks, and introduce a range of tools and resources available to help librarians effectively implement data curation, and particularly, preservation services.

This tutorial is being offered as part of the CRADLE (Curating Research Assets and Data Using Lifecycle Education) project, sponsored by the Institute of Museum and Library Services, under award #RE-06-13-0052-13.

Keywords

CRADLE

1. OUTLINE

Part I: Issues in Digital Data Curation and Preservation
A review of current issues around data curation including funding agency and publisher policies, the open access movement, and eScience trends that have made it necessary for library leaders to provide data curation services. Implications for preservation will be traced throughout the data curation lifecycle.

Part II: Research Data Curation Standards and Best Practices
An introduction to various data types and applicable standards and best practices for data curation with a focus on requirements for long-term preservation.

Part III: Data Management and Preservation Planning
In-depth exploration of tools and resources for archiving, preserving, and providing access to data in accordance with standards and best practices.

Part IV: Data Curation and Preservation Education
Additional exploration of data curation tools and services that are available to librarians and the researchers they support to assist them with data curation tasks and long-term preservation.

Upon completion of the tutorial, participants will be able to:

- Recognize different types of data and data curation and preservation issues specific to those types of data;
- Be knowledgeable of established and emerging standards and best practice for data curation and preservation;
- Understand funding agency and publisher policies, standard community practices, and other issues driving the need for digital data curation and long-term preservation;
- Understand the components of an effective Data Management Plan and the implications for long-term; and
- Be familiar with available tools and services to assist with data curation and preservation tasks.

2. TARGET AUDIENCE

Library practitioners working in libraries, archives, government agencies, corporations, or other organizations responsible for archiving, preserving, and providing access to research data. This tutorial is introductory; participants are not expected to have previous experience with the proposed topics.

3. NUMBER OF ATTENDEES

About 30.

4. FORMAT AND DURATION

Full-day tutorial

5. ORGANIZERS

Dr. Helen R. Tibbo, Alumni Distinguished Professor

School of Information and Library Science

tibbo@ils.unc.edu

Dr. Tibbo is an Alumni Distinguished Professor at the School of Information and Library Science (SILS) at the University of North Carolina at Chapel Hill (UNC-CH. She teaches in the areas of archives and records management, digital preservation and access, appraisal, and archival reference and outreach. She is also a Fellow of the Society of American Archivists (SAA) and was SAA President 2010-2011. She is the Principal Investigator of the IMLS-funded CRADLE project.

Thu-Mai Christian, Data Archivist

Odum Institute for Research in Social Science

tlchristian@unc.edu

Thu-Mai Christian has served as the Data Archivist for the Odum Institute since 2012. She is responsible for the day-to-day operations of the Odum Institute Data Archive as well as establishing and enforcing policies in accordance with archival standards and best practices. She received her Master of Science in Information Science degree with a concentration in Archives and Records Management from the School of Information and Library Science at the University of North Carolina at Chapel Hill. To continue her research in data management and preservation, she is also currently pursuing a Ph.D. in Information and Library Science on a part-time basis.

6. RELATED TUTORIALS

Curating Research Assets and Data Using Lifecycle Education

78th Annual Meeting of the Society of American Archivists August 11, 2014, 30 participants.

From Theory to Practice: Using ISO16363

Helen R. Tibbo
Sch. of Info. & Lib Science
201 Manning Hall CB#3360
UNC – Chapel Hill
+1-919-962-8063
Tibbo@ils.unc.edu

Nancy McGovern
Curation & Preservation Services
MIT Libraries
Massachusetts Institute of Technology
+1-617-253-5664
nancymcg@mit.edu

Barbara Sierman
KB National Library
of the Netherlands
Amsterdam, NL
+31 70 314 01 09
Barbara.Sierman@KB.NL

Courtney Mumma
Artefactual Systems, Inc.
Westminster, BC, Canada
+1 604-527-2056
Courtney@artifactual.com

Ingrid Dillo
Data Archiving & Networked Services
Den Haag, Netherlands
+31 6 12 16 69 89
ingrid.dillo@dans.knaw.nl

ABSTRACT

This tutorial will focus on an array of options and programs for audit and potential certification of trustworthy digital repositories. These will include self-audit, the European three-level model of certification, the Data Seal of Approval, peer-audit, ISO 16363 audit, and forthcoming certification of trustworthy repositories.

General Terms

Trustworthy repositories; Audit; Certification; Peer review; Self audit; External audit; ISO 16363

Keywords

Audit, management; Certification, Testing, and Licensing [The Computing Profession];

1. INTRODUCTION

Work toward assessing the quality of digital repositories can be traced to the early 2000's with the publication of OCLC's "Trusted Digital Repositories: Attributes and Responsibilites." This was based on the new OAIS ISO standard and was followed by the first version of TRAC (Trusted Repositories: Audit and Certification Checklist) in 2005 and a final TRAC version in 2007, both of which further developed the 2002 TDR document. In this same timeframe DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) appeared. In 2008, DANS created the Data Seal of Approval which has become the foundation of the European three-level framework for audit and certification.

Building upon OAIS, TRAC, DRAMBORA, and the DSA, the ISO 16363 standard is a formal framework for determining whether an organization is a Trustworthy Digital Repository. Published in 2012, the standard considers not only the technical infrastructure used for digital object management but also organizational infrastructure, and security risk management. In 2014, ISO 16919: Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories appeared.

In 2015, ANAB (the ANSI-ASQ National Accreditation Board) announced that it is developing an accreditation program for certification of organizations conforming with ISO 16363 Audit and Certification of Trustworthy Digital Repositories and ISO 16919.

Publications of these standards have led to numerous approaches to audit and potential certification. This one-day tutorial aims to de-mystify ISO16363 and other trustworthy repository best practices and provide attendees with insight into the practicalities of using these tools and approaches for formal audit or self-assessment.

2. INTENDED AND EXPECTED AUDIENCE

This tutorial is designed for those interested in improving their repository with information gained through some form of audit. These may be individuals with direct digital preservation responsibilities or those with strategic and oversight responsibility related to digital preservation.

Students who are planning for preservation careers should also find this session of interest.

3. OUTLINE OF TUTORIAL

This workshop will be composed of several sessions:

- Personal introductions and overview
- Overview of the notion of trustworthy repositories
- Introduction to various types of audit and forthcoming certification of repositories
- The European three-level model of certification and an exploration of Dutch initiative to get 5 major preservation organisations certified according to this model.
- Development and implementation of the Data Seal of Approval. The DSA was developed at DANS (Data Archiving and Networked Services) in the Netherlands.
- A brief history of the development of ISO 16363 and key elements of this standard.
- ISO 16363 training options.
- Self-assessment and peer review examples for ISO 16363 audits with an exploration of the DPM Management tool that is downloadable through Artefactual.
- Implementation and support of self- and peer-review audits for ISO 16363 through Archivematica. Examples for helping repositories with self-assessments and how Archivematica addresses specific requirements.
- Audit plan for a large federally-funded repository development project.
- Panel discussion of lessons learned from standards development and case studies.
- Small group discussion – participants will break into smaller groups to discuss implications for their own institutions and opportunities for collaboration.
- Reporting out – groups will report out about their discussions and the larger group will convey what they see as (individual and collective) next steps

4. OVERALL ORGANIZING STRATEGY

The tutorial organizers will advertise this session via various listservs and other social media.

5. WORKSHOP ORGANIZERS

- Helen R. Tibbo, Alumni Distinguished Professor, University of North Carolina at Chapel Hill
- Nancy Y. McGovern, Head, Curation and Preservation Services at MIT Libraries
- Barbara Sierman, Koninklijke Bibliotheek, National Library of the Netherlands, Research and Development Department
- Courtney Mumma, Consultant, US and International Community Development, Artefactual Systems
- Ingrid Dillo, Deputy Director at DANS - Data Archiving and Networked Services, The Hague, Netherland

Fedora 4 Tutorial

David Wilcox
DuraSpace
P.O. Box 138
Winchester, MA 01890
1-607-216-4548
dwilcox@duraspace.org

Andrew Woods
DuraSpace
P.O. Box 138
Winchester, MA 01890
1-607-216-4548
awoods@duraspace.org

ABSTRACT

Fedora is a flexible, extensible repository platform for the management and dissemination of digital content. Fedora 4, the newly released, revitalized version of Fedora, introduces a host of new features and functionality that both new and existing Fedora users are interested in learning about and experiencing first-hand.

This tutorial will provide an introduction to and overview of Fedora 4, with a focus on the latest features. Fedora 4 implements the W3C Linked Data Platform recommendation, so a section of the tutorial will be dedicated to a discussion about LDP and the implications for Fedora 4 and linked data. Fedora 4 is also designed to be integrated with other applications, so a section of the tutorial will review common applications and integrations patterns. Finally, attendees will participate in a hands-on session that will give them a chance to install, configure, and explore Fedora 4 by following step-by-step instructions.

General Terms

Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice; Training and education.

Keywords

Fedora, repository, linked data, open source.

1. OUTLINE

The tutorial will include four modules, each of which can be delivered in 1-2 hours.

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.

1.3 Hands-on with Fedora 4

It is quite simple to get up and running with Fedora 4. This module will give attendees an opportunity to experience Fedora 4 first-hand by following step-by-step instructions using a fully-functional virtual machine environment.

1.4 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 integration patterns, with a focus on some of the most popular application integrations.

2. DURATION

Full-day (6 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
- Gain familiarity with the Fedora 4 software
- Understand how to integrate Fedora 4 with external applications

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.

iPRES 2015

CHAPEL HILL – NOVEMBER 2–6

Proceedings of the 12th International Conference on Digital Preservation