19th International Conference on Digital Preservation 2023

Conference Proceedings

iPRES 2023 | Champaign-Urbana, Illinois

September 19-23, 2023

www.ipres2023.us

iPRES 2023

Digital Preservation in Disruptive Times

19th International Conference ■ Champaign-Urbana, Illinois ■ September 19–23, 2023

Proceedings of the 19th International Conference on Digital Preservation 2023

These proceedings contain the published and peer-reviewed submissions of the 19th International Conference on Digital Preservation. All other materials of the conference will be published on the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS): https://www.ideals.illinois.edu/units/541.

The IDEALS proceedings contain all submitted papers, panels, posters, workshops, and tutorials, as well as presenters' slides, optional additions, and collaborative notes taken during the conference.

The majority of the presentations at iPRES 2023 have been recorded and are now available on: https://mediaspace.illinois.edu/channel/iPRES+2023/320674362.

The conference photo are available on Flickr: https://www.flickr.com/photos/199245504@N05/.

Conference Organization

Program Committee

The success of iPRES 2023 was made possible due to the dedication and support of the following individuals.

General Chairs

Chris Prom, University of Illinois Urbana-Champaign
Tracy Seneca, University of Illinois at Chicago

Conference Organizing Committee

Tracy Seneca, University of Illinois Chicago
Chris Prom, University of Illinois Urbana-Champaign
Ruby Martinez, University of Illinois Urbana-Champaign
Brent West, University of Illinois System
Sara Bertier, University of Illinois Urbana-Champaign
Eden Irwin, University of Illinois Urbana-Champaign
J. Stephen Downie, University of Illinois Urbana-Champaign
Bethany Anderson, University of Illinois Urbana-Champaign
Kyle Rimkus, University of Illinois Urbana-Champaign
Joshua Henry, University of Illinois Urbana-Champaign
Dena Strong, University of Illinois Urbana-Champaign

Members of the organizing committee

Communications Committee

Ruby Martinez, University of Illinois Urbana-Champaign
Caitlin Perry, International Image Interoperability Framework
Angela Puggioni, Digital Preservation Coalition
Liselot Quisquarter, Flemish Institute of Archives

Papers and Panels

Helen Hockx-Yu, Notre Dame
Michael Nelson, Old Dominion
Neil Grindley, Jisc
Eld Zierau, Danish Royal Library
Michael Popham, Digital Preservation Coalition
Bethany Anderson, University of Illinois

Workshops and Tutorials

Tricia Patterson, Harvard University
Sandi Caldrone, University of Illinois
Seamus Ross, University of Toronto
Stephen Abrams, Harvard University
Villy Achieng Magero, Nairobi City County

Posters

Nathan Tallman, Penn State University
Carol Kussmann, University of Minnesota
Rachel Miller-Haughton, University of Illinois
Lauren Seroka, Library of Congress
Max Eckard, Bentley Historical Library, University of Michigan
Don Brower, University of Notre Dame

First Time Participant Committee

Kristen Allen Wilson, University of Illinois
Chelsea Denault, Michigan Digital Preservation Network
Emily Monks-Leeson, Library and Archives Canada
Juliet Awinja Erima, Moi University
Madeline Goebel, Library of Congress
Lauren Work, Yale University
Lance Stuchell, University of Michigan

iPRES 2023

Digital Preservation in Disruptive Times

19th International Conference ■ Champaign-Urbana, Illinois ■ September 19–23, 2023

Chicago Organizing Committee

Tracy Seneca, University of Illinois Chicago
Jeanne Long, Chicago Collections Consortium
Devin Savage, Illinois Institute of Technology
Ellen Keith, Chicago History Museum
Molly Szymanski, School of the Art Institute of Chicago

Remote Participation Committee

Joshua Henry, University of Illinois Urbana-Champaign
Dena Strong, University of Illinois Urbana-Champaign
Louise Curham, Charles Sturt University
Meghan Lyon, Library of Congress
Qianqian Yang, Sun Yat-sen University
Kyle Rimkus, University of Illinois Urbana-Champaign

Ad-Hoc Program Committee

Kate Murray, Library of Congress
Steven Gentry, Bentley Historical Library
Mikala Narlock, University of Minnesota
Winny Nekesa SKULLO, Public Procurement and Disposal of Public Assets Authority

Bake-Off Committee

Emily Shaw, Myriad
Micky Lindlar, TIB – Leibniz Information Centre for Science
Sibyl Schaefer, UC San Diego Library
Kate Flynn, University of Illinois, Chicago Library
Tracy Popp, University of Illinois Urbana-Champaign
Matthias Priem, Flemish Institute for Archives

Executive Committee

Tracy Seneca, University of Illinois Chicago
Chris Prom, University of Illinois Urbana-Champaign
Brent West, University of Illinois Urbana-Champaign
Ruby Martinez, University of Illinois Urbana-Champaign
Sara Berthier, University of Illinois Urbana-Champaign
William Kilbride, Digital Preservation Coalition
Dries Moreels, Ghent University Library

Reviewers

Stephen Abrams, Harvard University
Matthew Addis, Arkivum Ltd
Sawood Alam, Internet Archive
Gabriela Andaur, Archivo Nacional de Chile
Bethany Anderson, University of Illinois Urbana-Champaign
Thomas Baehr, TIB – Leibniz Information Centre for Science and Technology
Dara Baker, National Archives and Records Administration
Julianna Barerra-Gomez, Harvard Library
Heather Barnes, Wake Forest University
Charles Blair, University of Chicago Library
Karin Bredenberg, Kommunalförbundet Sydarkivera
Donald Brower, University of Notre Dame
Robert Buckley, University of Rochester
Birgitte Bullaert, Flemish Government
Carolyn Caizzi, Northwestern University
Sandi Caldrone, University of Illinois Urbana-Champaign
Bertrand Caron, Bibliothèque nationale de France
Yinlin Chen, Virginia Tech
Zijun Chen, National Science Library, Chinese Academy of Sciences
Euan Cochrane, Yale University Library
Elena Colon-Marrero, Bentley Historical Library, University of Michigan
Michael Day, The British Library
Mark Dehmlow, University of Notre Dame
Pilar Diazellis, National Archives of Chile
Dianne Dietrich, Cornell University
Max Eckard, Bentley Historical Library, University of Michigan
Elizabeth England, US National Archives and Records Administration
Juliet Erima, Moi University
Devora Geller, YIVO Institute for Jewish Research
Steven Gentry, Bentley Historical Library, University of Michigan
Andrea Goethals, National Library of New Zealand

iPRES 2023

Digital Preservation in Disruptive Times

19th International Conference ■ Champaign-Urbana, Illinois ■ September 19–23, 2023

Neil Grindley, JISC
Karen Hanson, Portico
Heikki Helin, CSC - IT Center for Science
Sarah Higgins, Aberystwyth University
Patrick Hochstenbach, Ghent University
Helen Hockx-Yu, University of Notre Dame
Darra Hofman, San Jose State University
Heidi Imker, University of Illinois Urbana-Champaign
Hellen Jerono, Moi University, Kenya National Archives
Shawn Jones, Los Alamos National Laboratory
Elizabeth Kata, International Atomic Energy Agency
Lauren Kata, New York University Abu Dhabi
Mat Kelly, Drexel University
Gladys Kemboi, WUR
William Kilbride, Digital Preservation Coalition
Alex Kinnaman, Virginia Tech
Martin Klein, Los Alamos National Laboratory
Carol Kussmann, University of Minnesota Libraries
Thomas Ledoux, Bibliotheque nationale de France
Michelle Lindlar, TIB - German National Library of Science and Technology
Lungile Luthuli, University of Zululand
Meghan Lyon, Library of Congress
Villy Magero, Nairobi City County Government
Basma Makhlouf-Shabou, Haute Ecole de Gestion
Ruby Martinez, University of Illinois Urbana-Champaign
Sharon McMeekin, Digital Preservation Consortium
Rachel Miller-Haughton, University of Illinois Urbana-Champaign
Jenny Mitcham, University of York
Emily Monks-Leeson, Library and Archives Canada
Dries Moreels, Ghent University
Kate Murray, Library of Congress
Kai Naumann, Landesarchiv Baden-Württemberg
Michael Nelson, Old Dominion University
Daniel Noonan, The Ohio State University
Alexander Nwala, William & Mary
Kieran O'Leary, National Library of Ireland
Jack O'Sullivan, Preservica Ltd
Trevor Owens, Library of Congress
Panagiotis Papageorgiou, University of Portsmouth
Julienne Pascoe, Library and Archives Canada
Tricia Patterson, Harvard Library
Stanislav Pejša, Purdue University
Maureen Pennock, British Library
Michael Popham, Digital Preservation Coalition
Chris Prom, University of Illinois Urbana-Champaign
Andreas Rauber, Vienna University of Technology
Seamus Ross, Faculty of Information, UofToronto
Robin Ruggaber, University of Virginia Library
Sibyl Schaefer, University of California San Diego
Lynda Schmitz Fuhrig, Smithsonian Institution Archives
Tracy Seneca, University of Illinois at Chicago
Lauren Seroka, Library of Congress
Michael Shallcross, Bentley Historical Library, University of Michigan
Arif Shoan, Qatar National Library
Tobias Steinke, Deutsche Nationalbibliothek
Paul Stokes, Jisc
Nathan Tallman, The Pennsylvania State University
Heather Tompkins, Library and Archives Canada
Yvonne Tunnat, Leibniz Information Centre for Economics (ZBW)
Susanne van den Eijkel, Koninklijke Bibliotheek (National Library of the Netherlands)
Remco van Veenendaal, Nationaal Archief
Matthias Vandermaesen, Ghent University
Alexandra Vidal, independent
Anna Vögeli, University Library Bern
Michele Weigle, Old Dominion University
Jess Whyte, University of Toronto
Lotte Wijsman, National Archives of the Netherlands
Qianqian Yang, Sun Yat-sen University
Eld Zierau, Royal Danish Library

iPRES 2023

Digital Preservation in Disruptive Times

19th International Conference ■ Champaign-Urbana, Illinois ■ September 19–23, 2023

Sponsors

iPRES 2023 could not have been possible without the support from the following organizations who sponsored the conference.

Preservica

libnova

DIGITAL BEDROCK
Keeping Your Digital Assets Safe and Evergreen

infotel

aws

apryse

AM Primary sources reimagined

artefactual systems inc.

arkivum

DigitalPreservationCoalition

nestor

PORTICO

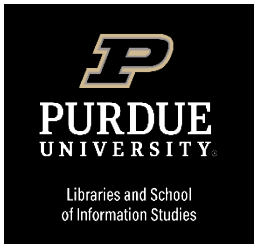
docbyte

Council on Library and Information Resources

iPRES 2023

Digital Preservation in Disruptive Times

19th International Conference ■ Champaign-Urbana, Illinois ■ September 19–23, 2023



iPRES 2023

Digital Preservation in Disruptive Times

19th International Conference ■ Champaign-Urbana, Illinois ■ September 19–23, 2023

Conference Summary

The University of Illinois at Urbana Champaign is proud to release the iPRES 2023 proceedings on World Digital Preservation Day. We hope you find them profitable and useful for your work.

Doing so is a fitting capstone to the community effort that underpins digital preservation work so evident during the iPRES 2023 meeting, held from September 19 - 22, 2023. As conference co-chair, along with Tracy Seneca, I can truly say that organizing the conference was THE highlight of my professional career to date. The sense of welcome, collegiality, rigor, and fun that the entire conference planning team—our program committee, local organizers, peer reviewers, and supporters— brought to the table, was truly extraordinary and inspiring!

If our conference theme—Digital Preservation in Disruptive Times—implied that our field was seeing some stress, this was far from evident, judging by the sense of camaraderie we exemplified in our interactions, both on site in Champaign-Urbana, Illinois, and online in our hybrid environment.

As you will see when you read the Proceedings, our contributors provided many new and original insights, around our five conference themes (All Together Now, Digital Accessibility, Inclusion and Diversity, From Theory to Practice, Immersive Information and Sustainability).

Delegates at the opening ceremony

I hope you will read the proceedings cover to cover, although we certainly understand that you can also dip into just those papers of interest. For this reason, we are providing both a PDF of the entire proceedings and access to individual articles, through the University of Illinois Institutional Repository, IDEALS. Videos of many sessions are also available here.

I am eager to point out that Illinois could not—indeed did not—do this alone. We were fortunate to build on the excellent work and leadership provided by the iPRES Steering Committee, as well as the excellent example set by our Scottish predecessors, and that over 300 on site registrants and over 200 online registrants attended iPRES 2023. We were perhaps most grateful that 14 scholarship recipients attended in person, and three remotely. In addition, we were able to waive registration fees entirely for all attendees who indicated that they were self funded. This was all due to the generosity of our sponsors and academic partners, noted elsewhere in the proceedings. They contributed over $130,000 of direct financial support to the iPRES community.

On a personal note, I am proud to say that iPRES gives a feeling of home. I love being part of this community. It spans many areas of practice, and each year brings both a set of familiar and new faces.

As I look forward to iPRES 2024 in Ghent, I hope you will also join me in raising a virtual glass to the role that iPRES plays in fostering an expansive and expanding community of international experts dedicated to celebrating all things digital preservation!

-Chris Prom, General Co-Chair

Keynotes

Sherry Williams

Sherry Williams was born and raised on the south side of Chicago in the Englewood Community. She is Founder and President of the Bronzeville / Black Chicagoan Historical Society and a graduate of the University of Illinois School of Information Science.

Williams led African American cultural programs at the Pullman State Historic Site on the Senator Stephen A. Douglas Tomb Site and Monument Park grounds from 2007-2017. In 2009, Williams was voted Vice President of the Pullman Civic Organization. She also served as a board member of the Chicago Cultural Alliance, from 2000-2005.

She is an active member of the Afro American Genealogical and Historical Society (Chicago Chapter); a board member of the Bronzeville Trail Task Force, Inc.; a board member of Chicago Coalition of Park Advisory Councils; an advisory member of Illinois State Historical Society; a former commissioner of the Amistad Commission of the State of Illinois (2010-2012); an institutional member of the Chicago Cultural Alliance; a partner

institution of Choose Chicago; and a board secretary of the Burnham Park Advisory Council.

Williams was the opening keynote for iPRES 2023, "Sherry Williams: A Conversation with Sherry Williams about Community-focused Digital Preservation". The recording of the presentation " is available for viewing here: https://mediaspace.illinois.edu/media/t/1_54n5155o/320674362.

Dr. Ricardo Punzalan

Dr. Ricardo L. Punzalan, associate professor at the University of Michigan School of Information, is a scholar of archives and digital curation. He studies community access and use of anthropological data in archives, as well as the digitization of ethnographic records held in libraries, archives, and museums. His research has established and shaped practices of virtual reunification and digital repatriation of cultural heritage collections. To do this work, he designs and carries out community-based, participatory research projects, which incorporate the perspectives of cultural heritage stakeholders beyond academic researchers. His scholarship has brought to the fore the critical challenges faced by underserved and Indigenous communities and has created dialogs between communities and cultural institutions. He

Keynote speakers Quinn Dombrowski, Sherry Williams, and Dr. Ricardo Punzalan with Chris Prom

co-directs ReConnect/ReCollect: Reparative Connections to Philippine Collections at the University of Michigan, a project that develops the framework for, and the practice of, reparative work for Philippine collections acquired by the university during the US colonial period. He is currently co-chair of the Archival Repatriation Committee of the Society of American Archivists and on the Board of Trustees of the Library of Congress American Folklife Center. He was recently inducted as a Fellow of the Society of American Archivists.

Punzalan was the second keynote and presented "Reciprocity, Reparative Actions, and Decolonial Work". The recording of the presentation is available for viewing here: https://mediaspace.illinois.edu/media/t/1_02alt5dq/320674362.

Quinn Dombrowski

Quinn Dombrowski is the Academic Technology Specialist in the Division of Literatures, Cultures, and Languages, and in the Library, at Stanford University. Prior to coming to Stanford in 2018, Quinn's many digital humanities adventures included supporting the high-performance computing cluster at UC Berkeley, running a tool directory with support from the Mellon Foundation, writing books on Drupal for Humanists and University of Chicago library graffiti, and working on the program staff of Project Bamboo, a failed digital humanities cyberinfrastructure initiative.

Quinn has a BA/MA in Slavic Linguistics from the University of Chicago, and an MLIS from the University of Illinois at Urbana-Champaign. Since coming to Stanford, Quinn has supported numerous non-English digital humanities projects, started a Textile Makerspace, developed a tabletop roleplaying game to teach DH project management, explored trends in multilingual fanfic, and started the Data-Sitters Club, a feminist digital humanities pedagogy and research group focused on Ann M. Martin's 90's girls series "The Baby-Sitters Club". A co-founder of Saving Ukrainian Cultural Heritage Online (SUCHO), Quinn has been working to preserve and augment Ukrainian digital cultural heritage since Russia's invasion in February 2022. Quinn also serves as co-President of the Association for Computers and the Humanities, the US-based professional association for digital humanities.

Dombrowski closed the conference with "Takes All Kinds: Grassroots Digital Preservation in a Crisis and Beyond." The recording of the presentation is available for viewing here: https://mediaspace.illinois.edu/media/t/1_zw8wby37/320674362.

Peer Reviewed Program

The conference program included sessions of paper presentations, panels, posters and bake-off demonstrations, preceded by workshops and tutorials.

The conference program consisted of up to five concurrent sessions each day. Mostly all sessions were streamed live and made available to online delegates, apart from a portion of the ad-hoc program. The recordings were available to delegates on the platform and will be available under mostly open access thereafter.

Tuesday involved Workshops and Tutorials with a welcome reception at Memorial Stadium. Tuesday opened with a keynote speaker followed by concurrent strands in sessions of 90 minutes. On Wednesday and Thursday, the order was reversed with the keynotes happening at the end of the day. A session was dedicated to Posters on Thursday; and ad-hoc activities including Games and the Bake Off carried on throughout the conference.

Following a peer review process iPRES 2023 accepted a total of 77 submissions (a breakdown is provided below).

Contribution Type	Number of Submissions
Long Papers	15
Short Papers	30
Panels	11
Workshops & Tutorials	12
Posters	9

Ad-Hoc Program

In addition to the peer reviewed program, iPRES 2023 also had a non-peer reviewed program which included lightning talks, games, and birds of a feather sessions taking place throughout the conference. Following a lightweight review process the Ad-Hoc program accepted 72 submissions (a breakdown is provided below).

Contribution Type	Number of Submissions
Bake-Off	11
Lightning Talks	31
Games	6
Birds of a Feather	7
Virtual Site Visits	17

Social Program

On Tuesday evening, the University of Illinois was pleased to host a welcome reception, sharing the flavor of a BigTen University campus with our delegates, with hors d'oeuvres and drinks served at the Colonnades Club, in the University of Illinois Memorial Stadium. University Provost John Coleman welcomed us to campus and was presented with a gift: A Champaign Urbana farmers hat. And the group was treated to a performance by the Marching Illini. Resonant tones from the brass section and drums filled the stadium, while the fall light filtered in the stadium, making for truly impressive video opportunities, as well as numerous selfies.

On Wednesday, all delegates were invited to the conference dinner, which is a special iPRES tradition. Taking a short bus ride,

delegates arrived at a local event venue, Carmon's. After being handed a special treat from an antique popcorn car, delegates entered the venue: A former train shed, renovated from the days when Interurban streetcars connected the twin cities to outlying areas. With apple time for socializing and drinks (sponsored by Clarivate Ex Libris, delegates also enjoyed a buffet-style dinner. The evening concluded with a special performance by 90's Daughter, Champaign-Urbana's premier cover band. Delegates shed the inhibitions (within limits) on the dance floor, and everyone was happy for the opportunity to connect informally.

The First Time Participants Committee also arranged for social dinners around the Champaign Urbana area on Thursday evening, as well as a Daily Morning Rundown to prepare and connect first time attendees ahead of each conference day.

Finally, on Friday, delegates were treated to the first of Chicago's events, which included an evening reception at the University of Chicago's Gleacher Center. On Saturday, September 23, several of Chicago's cultural heritage organizations welcomed iPRES attendees for guided tours led by local staff. Additionally, there was an exclusive tour led by keynote speaker, Sherry Williams.

Delegates with keynote speaker Sherry Williams at the Ida B. Well Monument in Chicago, Illinois

While an oft-mentioned tethered hot air balloon never made its appearance, due to unfavorable atmospheric conditions, the conference hosts made the best of this situation, with a running joke.

We were proud to uphold the fine traditions of iPRES social events, which are so critical in building and reinforcing the relationships that bind us in confraternity. We all look forward to future iPRES social events, for this very reason.

iPRES 2023 Awards

Following the iPRES conference tradition, iPRES 2023 took the opportunity to recognize outstanding contributors to celebrate these in a set of conference prizes.

This year there were five prizes awarded for: Best Paper, Best Paper, Best First Time Contribution, the Angela Dappert Memorial Award, and the Co-Chair's Award for Outstanding Overall Contribution to the Conference.

Best Paper of iPRES 2023 sponsored by Nestor

The Best Paper Prize was awarded to 'Long-Term Preservation of a Software Execution State' by Rafael Gieschke, Klaus Rechert, and Euan Cochrane.

Best Poster of iPRES 2023 sponsored by the University of Iowa

The Best Poster Prize went to: 'Embedding Preservability: Iframes in Complex Scholarly Publications' by Karen Hanson, Jonathan Greenberg, Thib Guicherd-Callin, Scott Witmer and Angela T. Spinazzè.

Best First Time Contribution sponsored by the Digital Preservation Coalition

The iPRES 2023 Best First Time Contribution was awarded to Maurren Kenga for her poster "Digital Accessibility, Inclusion and Diversity: Digitization of Indigenous Agricultural Knowledge in Shaping Food Security Across the Kenyan Coastal Region" and for her other contributions to the conference as a commentator, inculcator, and participant at the multiple sessions she attended.

The Angela Dappert Memorial Prize sponsored by Adam Farquhar

The Angela Dappert Memorial Prize was awarded to Meghan Lyon and Grace Bicho for their paper, "Emerging Quality Assurance Practices in the Library of Congress Web Archives".

Co-Chair's Award for Outstanding Overall Contribution to the Conference

This year's iPRES included a special award, which was presented to Marion Ville for her paper, '2013 - 2023: A Review of Ten Years of Email Archiving in France.'

iPRES 2023

Digital Preservation in Disruptive Times

19th International Conference ■ Champaign-Urbana, Illinois ■ September 19–23, 2023

iPRES 2023 Submissions

2013 - 2023: A Review Of Ten Years Of Email Archiving In France

Marion Ville
Ministry for Europe and Foreign Affairs
Vitam Program
France
marion.ville@culture.gouv.fr

Abstract – Emails are an essential medium of communication. Their management is an organizational, security, legal and financial issue for all organizations.

The interdepartmental digital archiving Vitam program, which develops a digital archives management system on behalf of the French government, could not help but wonder about the acquisition, preservation and access to the documents and data contained in email archives. In 2013, it carried out a proof of concept on email archiving, at a time when few archive services in France had embarked on the acquisition of this type of document.

Ten years later, the French landscape in terms of digital recordkeeping has changed significantly. In practice, some archives have put in place strategies for archiving email and tools have been made available to assist in their effort.

This article looks at the transition from theory to practice of a more operational acquisition of this type of archive in a French context.

Keywords – Email archiving, Appraisal, Preserving email, Tools, Proof of concept

Conference Topics – From Theory to Practice.

I. Introduction

Email archiving is now a common and unavoidable practice in French central administration.

In 2013, when the interdepartmental digital archiving Vitam program, which develops the digital archives management system called Vitam on behalf of the French government and has currently around 60 partner institutions from the public and private sectors, decided to carry out an international and national state-of-the-art study, as well as a proof of concept on the subject, there was little national expertise.

The 2013 report that was produced was a first milestone in the knowledge and mastery of the issues surrounding this type of archive [1].

Since then, it has been noted that French public archive services, in charge of record management and long-term archiving, have put in place strategies and methods for email acquisition. Tools have also been made available to facilitate the process.

These new practices and choices have brought to light new challenges facing the mass of email acquired, in terms of complexity of the processing to be carried out, preservation and access to email.

II. A Vitam Proof Of Concept

A. Objective and Process

The Vitam project team, in partnership with the Ministries of Culture, Foreign Affairs and Defense, conducted a proof of concept on email archiving between March and June 2013. The main objectives of this study were to:

- define a strategy for email acquisition, processing and preservation, adapted to the different contexts of the Vitam project partners,

- identify the tools and technical functionalities required for the technical processing of these emails, prior to their transfer in the Vitam software,

- define a metadata model for the email archives in accordance with the requirements of the Vitam project team.

For this purpose, the Vitam project team based itself on a review of the literature and existing

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

national and international experiences. The team also carried out technical tests. A report on this work was published in October 2013 [1].

B. First Part: Literature Review

The first part of the study was to produce a summary in French of the work carried out at both national and international levels.

In France, several documents containing technical and management recommendations were published in 2008-2009. On 3 June 2009, the French Archives Department (DAF) published an instruction disseminating and commenting on the directive published by the State Archives of Belgium [2]. Associations such as the Archival Policy and Project Managers Club (CR2PA) [3], FeDISA [4] or the Association of French Archivists had produced white papers or advice sheets. Institutions such as the National Archives [5] or the National Library of France had produced guides for internal distribution.

At the international level, at that time, the first summaries of experiments were produced [6]. As a result of this literature review, the members of the Vitam project team concluded that three approaches to email archiving existed:

- a "pedagogical" approach, the most widespread, whereby actors in the digital and archiving world tried to alert users to the consequences of their use of the messaging tool and tried to provide them with advice and guides to good practice (State Archives of Belgium [7], archives of the States of Alabama, South Carolina or Texas, Bibliothèque et Archives nationales du Québec (BAnQ) [8], The National Archives (TNA) [9], etc.);

- an "acquisition and preservation" approach, with a few major projects (the DAVID project of the Antwerp Municipal Archives [10], the project of the National Archives of the Netherlands [6], the archiving policy of the National Archives of Australia [11], the Collaborative Electronic Record Project led by the Smithsonian Institution Archives and the Rockefeller Archive Center [12]), which aimed to respond to the problems of medium-term and long-term acquisition and preservation of emails, in particular through the design of technical tools;

- a diplomatic approach, based mainly on integrity issues via the renewal of diplomatics initiated by the University of British Columbia and in particular by Professor Luciana Duranti as part of the InterPARES group [13], and addressed by the English project InSPECT [14].

C. Second Part: Experimentation

The second part of the study aimed to test the conversion from one format to another (.eml to .mbox, .pdf; .mbox to .csv, .eml, .pdf; .pst to .eml, .mbox) and the extraction of messages and attachments.

These tests were carried out jointly by the Ministry of Defense (Defense Historical Service-SHD) and the Ministry of Culture and Communication (National Archives and IT Department). They were carried out on the most commonly used email clients in the context of the French public administration: Thunderbird and Outlook.

They were performed with:

- two software packages on the market, Aid4Mail and EmailChemy, which ensured the conversion of messages and emails into standard formats (.mbox, .eml, .pdf, .csv);

- CERP Email Parser and Xena, software developed by archive services, which provided email processing in order to generate an XML envelope;

- tools available in open source libraries, Apache POI, CLibPST, Java LibPST, Java Mail, Mime4J, Tika, which provided metadata extraction from messages and emails, as well as processing to extract headers, bodies and attachments and to identify file formats, and DROID to identify file formats.

D. Conclusions and Proposals

1) For the authors of the study, the real challenges in preserving emails did not appear to be technical, but organizational, legal and archival, particularly for personal emails. The Vitam project team concluded with a few proposals:

- Organizational proposals: it was considered necessary to point out, at the highest level of the organizations, some simple rules relating to the use of email accounts and the nature of the information exchanged. Awareness-raising actions could be envisaged.

- Legal proposals: it seemed useful to draft standard clauses in internal regulations and IT policies, as well as to draw up a standard protocol to be signed by staff members when capturing their messages.

- Technical proposals: The transmission to IT departments of instructions based on the recommendations of the InterPARES working group could be a first step. The development of additional plug-ins, particularly for the Thunderbird client, which would allow important messages to be exported on the same principle as the LiveLink project run by the Republic and Canton of Geneva, was a second interesting approach.

- Archival proposals: it seemed important to define a strategy for email acquisition adapted to the uses of the various categories of administrations and producers, by identifying the target email accounts to be acquired in each organization, a suggestion made by the French Archives Department in 2009. Finally, the experimentation of semantic analysis tools was also considered an interesting line of thought. The development of an XML schema for representing email archives could constitute a first action that the archive administration could initiate. Finally, the diffusion in French of the work carried out at the international level seemed to be useful, as well as partnerships with other archive services at the international level.

2) In addition, following the literature review and the tests carried out, and taking into account the state of its thinking in terms of metadata modelling, the Vitam project team began to define a methodology for collecting email accounts and messages.

- Concerning the acquisition, the tests carried out led the project team to recommend extractions of emails in .eml, .mbox and .pst formats. For accounts exported in .pst format, it was recommended to reduce as much as possible the processing time between the export and the transfer to the competent public archive service, in order to avoid problems linked to the fast obsolescence of the format.

- Concerning the constitution of SIPs, in a platform based on the Vitam software, the authors of the study proposed that the Submission Information Packages (SIPs) corresponding to email accounts should take the form of a zipped file including an XML format file describing the email archive and complying with the Standard d'Echanges de Données pour l'Archivage (SEDA) [15] and the hierarchical structure of the folders, with, for each message, the body of the message, the associated attachments, and the messages in their original format.

III. From Theory To Practice

Since 2013, new studies and methods for email archiving have been developed both internationally [16] [17] [18] and nationally, where, based on these various work, procedures have been established by certain French archive services, whether they are in charge of records management (Mission Archives of ministries, Council of State, etc.), long-term archiving (National Archives) or both (Ministry for Europe and Foreign Affairs), and tools have been made available.

A. Building Archiving Strategies

1) Identifying The Emails to Acquire

French public archives, following the CAPSTONE approach [19], have chosen to select not all of the emails produced, but the email accounts deemed to be the most engaging. For example, the Ministry for Europe and Foreign Affairs has identified around 500 email accounts amongst the 15,000 or so managed by the Ministry. The email accounts selected include those of deputy directors or assistant directors [20]. Other ministries and government structures follow the same principle. For example, the Mission Archives of the Prime Minister's Office has targeted the email accounts of the Prime Minister's direct collaborators and, within the administration, the email accounts of the most senior civil servants (e.g. director of the National Institute of Public Service, Secretary General of the Government, director of the Digital Department, etc.). For intermediate-level emails, the Mission's archivists have also prescribed only strategic emails to be archived [21].

Most of the time, the acquisition is carried out when the email account is closed. Since 2012, the Mission Archives attached to ministries have been acquiring messages when ministers and their staff have left office, especially after elections. The Mission Archives of the Prime Minister's Office has already collected no less than 180 email accounts since then. And this type of acquisition has been increasing: while only two Mission Archives had acquired 7 messaging accounts in 2012, at least nine Mission Archives have gathered no less than 577 accounts in 2022.

A more targeted acquisition of email archives can also be carried out. This involves acquiring email accounts in use to carry out a specific mission, for example the organization of a particular event. Thus, the Ministry for Europe and Foreign Affairs chose to acquire the email accounts of the organizers of the COP21, the 2015 Paris climate change conference. The collection has included 1,772 email accounts, in various formats (.pst, .msg, .msf, .eml) [22]. Some Mission Archives did the same for the archives produced during the COVID-19 pandemic, although the acquisition strategies were different, depending on the circumstances and their organizational context. The Mission Archives of Social Ministries initially set up a procedure for acquiring the email accounts of agents employed as reinforcements in the health crisis center. 143 e-mail accounts were collected in this context.

However, is it appropriate to keep all of these email accounts? Some of them contain only a few records. The Mission's archivists are already planning a reappraisal of these email archives. The Archives Department of the Ministry of Justice proceeded differently. The archivists interviewed staff members of the health crisis unit in order to appraise their documentary practices and to determine the emails to acquire. When the producer had not prepared a folder on the subject, it was decided to acquire only messages filtered through approximately thirty keywords established with the agents' agreement. This has represented 100 GB of data, over a period from January 2020 to December 2022, for 15 personal email accounts and 3 institutional email accounts.

2) Acquisition Method

There are some differences in acquisition strategies. Mission Archives prefer to acquire email accounts in their original format, mostly in .pst format. Archivists sort the records before or at the time of acquisition, sometimes with the help of the producer. Afterwards, they do not process them any further before transferring them to the National Archives for long-term archiving, in order to preserve the email archives' integrity.

The National Archives, like the Ministry for Europe and Foreign Affairs, have chosen to acquire not only the raw export as obtained from the producer but also a version processed according to a specific protocol aimed at extracting messages from the .pst or .mbox container [23]. The National Archives accept raw exports in .pst and .mbox formats. This makes it possible to create a first SIP. The Ministry for Europe and Foreign Affairs accepts more formats [22]. In addition to this export, the National Archives require that messages in the form of .eml and .txt files, as well as attachments linked to the messages, are extracted from the .pst and .mbox containers. This constitutes a second SIP. This extraction is performed with the ReSIP tool provided by the Vitam program [24]. This tool can also be used for other processing: unzipping attachments, deleting folders without messages, folders containing messages that are considered private, business cards in .vcf format, logos in .emz format and files with the .dat extension [25]. It is this export that is meant to be used in case of an access request. This is also the way in which these two institutions guarantee the durability of this type of archive.

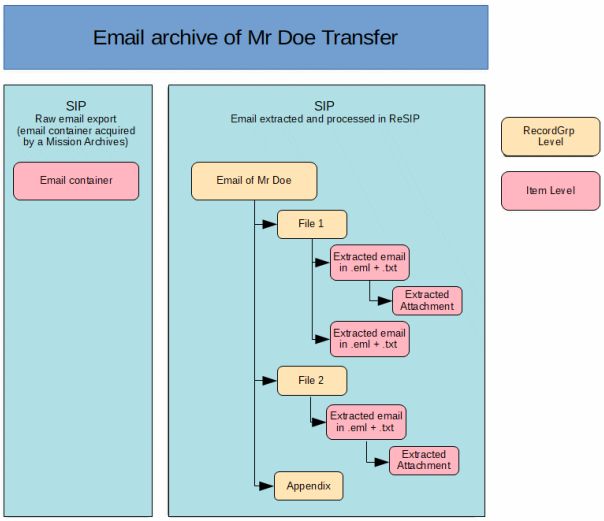


Fig. 1. Diagram of an email archive transferred to the National Archives

B. Strategies Formalization and Legal Construction

In the last few years, central government archives have begun to formalize their strategies in internal procedure notes and documents. They were made possible by the work on email archiving and digital preservation undertaken by the Vitam program [1], continued by the Interministerial Service of the French Archives, which has published a *vade mecum* on the subject [26], as well as a note from the Interministerial Delegate for the French Archives dated 18 May 2020, which has requested central administration services to adopt an email archiving

policy [27]. The formalization of these strategies has also been encouraged by exchanges between institutions within the Vitam program and within the National Formats Watch Unit. The latter was formed in 2019 to create a national space for thinking about and exchanging information on the issue of file formats and includes currently more than ten public institutions [28]. Internal experiments and studies have also confirmed these strategies, as was the case at the Ministry for Europe and Foreign Affairs [22] and the Council of State [29].

These strategies appear both in guidelines on the use of email accounts [30] and in notes emanating from the institution and aimed at raising awareness among departments about email archiving or informing them about acquisition procedures [29]. They take the form of lists of email accounts to be acquired, methodological guides designed to facilitate the appraisal of messages [21], practical sheets explaining how to create SIPs in conformity with the expectations of the archives [25][29] or defining roles and responsibilities, internal procedures aimed at organizing the acquisition, transfer or even search, valid for the whole institution or joint with certain departments. Very often, these documents are for internal use and are not disseminated outside the institutions concerned. It would be interesting to make them consistent.

C. Proposal of A Model

The Vitam program has also designed a model for structuring email accounts. It was first implemented with the MailExtract library [31], which has been integrated since 2019 into the Sedatools library, which is itself used by the ReSIP tool [24]. The library offers the possibility to import email accounts and messages in .pst, .msg, .eml and .mbox formats and to process them, offering several functionalities:

- extract messages from a .pst or .mbox container and migrate them to other formats: .eml, .txt,

- extract agendas and contact lists as a spreadsheet in .csv format,

- extracting the metadata of email accounts, and the textual content of the body of messages and their attachments as metadata embedded in a SEDA compliant XML file [15] or as a spreadsheet in .csv format.

The metadata extracted to this point from emails are:

- for folders, a description level corresponding to a group of documents (RecordGrp), their title, as well as the dates of the oldest and the most recent message;

- for each message, a description level corresponding to an item (File), its subject, the original identifier corresponding to the identifier of the message, the sender, the recipients and the addressees, the dates of expedition and reception, the reference to another message and the body of the message;

- for each extracted attachment, a level of description corresponding to an attachment (File), the name of the file, as well as a description specifying that it is an attachment (cf fig.1) [32].

IV. Current Challenges

A. A Documentary Mass to Process

In the presence of so many messages, it is difficult to quickly identify which messages should be deleted, even if archivists know which types of emails should be deleted (as personal messages, mailing lists, etc.). Some archives search for pre-defined keyword lists using ReSIP or Outlook [29]. However, this is a tedious and incomplete process.

There is a need to navigate in depth through email accounts in order to facilitate their processing. To address this need, the Mission Archives of Social Ministries has developed the Archifiltre-Mails tool [33]. The first version, released in the autumn of 2022 allows users to:

- import messages from Outlook (except for Office 365, which it is planned to support soon);

- explore emails and view messages by email domain, contacts by domain, years by contact, and then the messages themselves;

- export messages in .eml format;

- extract metadata and message content in spreadsheet form in .csv, .xlsx or .json formats. Date, sender, recipients, subject, message content, path in the classification plan, number, name and size of attachments are retrieved;

- add "delete" or "keep" tags;

- obtain statistical information.

The primary purpose of the tool, which is still being developed, is to be able to quickly identify messages that can be deleted [34].

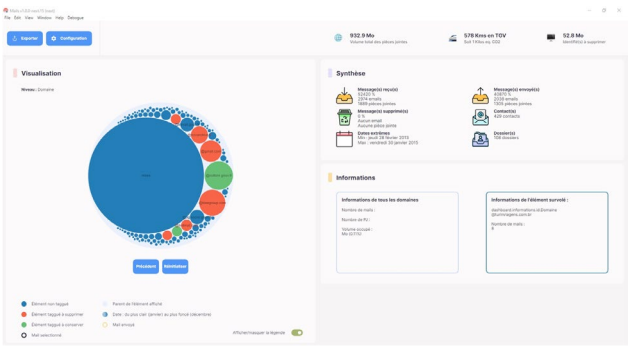


Fig. 2. Archifiltre-Mails dashboard and data visualisation.

The processing of messages may also be difficult or impossible without the original client, due to several factors:

- compatibility problems between the archive format and the software used to perform the processing (especially Outlook);

- digital workplaces and tools that are not sufficiently powerful to carry out the required processing, due to the size of the email archive;

- the errors contained in the email archives themselves. These are generated by the original client or by the producer. They are mainly detected during import attempts in existing tools such as ReSIP or Aid4Mail. They may be generated by duplicate messages, corrupted emails archives, attachments whose names exceed the limits allowed by Windows, whose encoding is not recognized, whose upload is blocked by another tool or which have been deleted. Repair is then manual and can be time consuming [22].

All these factors are often not well known by the profession. There is a need to increase competence in the issues related to the processing of messages.

B. A Documentary Mass to Preserve

E-mail accounts represent a large part of the digital archives acquired in recent years by the central administration's archive services. At the Ministry for Europe and Foreign Affairs, they constitute more than 40% of the stock preserved in their digital repository. The same can be said of the National Archives, where messages extracted from ReSIP represent about half of the descriptive and technical metadata recorded in the database for less than 0.5% of the deposits. If the flow of email acquisition continues, combined with a policy of extracting messages from ReSIP, this may raise a problem of technical maintainability for the repository system. This technical issue only applies for archives that have chosen to extract messages, a choice that is justified from an archival point of view, as it offers a guarantee of access and durability for this type of archive.

In order to reduce the number of extracted files and metadata, the Vitam program has proposed to test an experimental mode of compacting the extracted messages in version 2.7 of ReSIP [22]. Rather than having as many levels of description as there are folders and messages, the idea is to group all or part of the metadata extracted from the email archive at the most relevant level, determined by the archivists, in order to maintain the possibility of searching in the messages. The converted folders and messages, as well as the attachments, are placed in a .zip file.

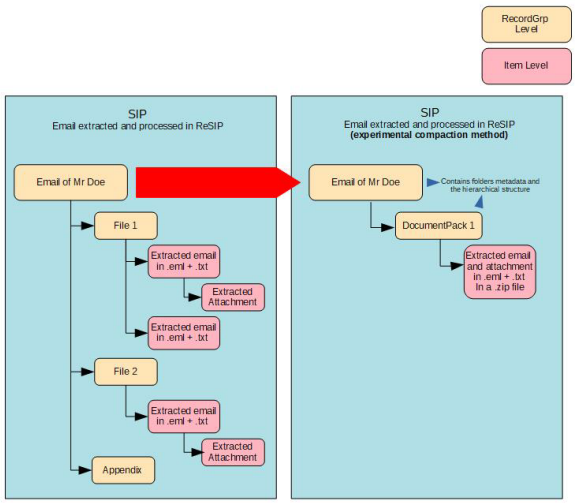


Fig. 3. Comparison between standard extraction and experimental extraction with ReSIP

The version extracted from the emails would then be smaller in terms of the number of levels of description, as well as the number and size of the files. Nevertheless, it raises questions in terms of:

- Durability, as the .eml and .txt files and attachments are encapsulated in a .zip file, which is not a good candidate for digital preservation.

- Access. How do you make the link between an email identified in the metadata and the file encapsulated in the .zip file?

The next work to be undertaken by the Vitam program will aim to answer these various questions by proposing functionalities for accessing these .zip files, if this experimental mode is proven successful.

C. A Documentary Mass to Access

New types of requests for access, formulated under the Code of Relations between the Public and the Administration (CADA law), are emerging [35]. These queries differ from the usual requests since they concern very broad themes, or even the occurrence of a word. They require searches on a large number of recent and unclassified sources, such as email archives. However, each of the current archiving scenarios does not completely solve the problems related to the access of this type of archive.

In the case of the Mission Archives that acquire emails in a container format, access is done manually by reimporting them into Outlook or ReSIP, or even into Archifiltre-Mails. But the operation is not easy, because it takes time if the email archive is voluminous and the import may not be successful because of its volume, the archive itself or the archivist's digital workplace. One of the solutions envisaged to facilitate future searches would be, at the time of email acquisition, to generate a .csv file from ReSIP or Archifiltre-Mails, which would include the metadata of the messages and would be archived at the same time as the email, because this format is easy to use for consultation. After these issues related to import, archivists must understand the organizational logic of the email, appraise the messages found in regard to the respect of privacy and rights of communicability and extract those that have been retained for access. This method is currently impractical and time-consuming.

The National Archives, for their part, are currently studying other ways of accessing email accounts and messages in order to facilitate their consultation. They are currently studying Ratom [36], but also epadd++ [37] and Pêle-mél [38]. The latter tool is developed by the Pêle-mél project team, composed of teacher-researchers from the University of Angers, and funded by the Ministry of Culture. It is a prototype for exploring and visualizing acquired email accounts, using extracted messages in .eml format only. It aims to facilitate access to acquired email accounts by using artificial intelligence and machine learning technologies [38].

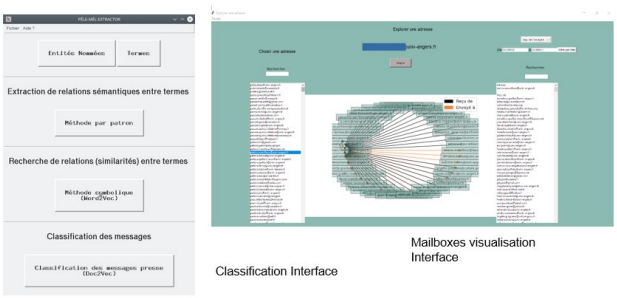


Fig. 4. Pêle-mél dashboard and data visualisation.

V. CONCLUSION

The Vitam program has provided a solid initial basis for further discussions for public archives thanks to its proof of concept, and then tools to facilitate email archiving. From a theoretical basis, we note that, in France, we have moved on to a more operational stage for this type of acquisition.

Nevertheless, in the face of massive documentary volume, this same phase has led to new questions, in terms of processing, preservation and access, as well as new experiments. Email archiving has not yet finished making waves

AKNOWLEDGMENTS

The author would like to thank Edouard Vasseur (Ecole nationale des Chartes), as well as Emeline Levasseur (National Archives), Camille Tatger (Ministry for Europe and Foreign Affairs), Nathalie Morin (Mission Archives of Prime minister's Office), Anne Lambert (Mission Archives of Social Ministries), Sarah Harroche (Archives Department of the Ministry of Justice).

REFERENCES

[1] Vitam, L'archivage des messageries électroniques. Preuve de concept VITAM. Paris: Ministère de la Culture et de la Communication, Ministère de la Défense, Ministère des Affaires étrangères et du Développement international, October 2013

[2] Direction des Archives de France, Instruction DITN/RES/2009/007 du 3 juin 2009. Paris: June 2009

[3] Club des responsables de politiques et projets d'archivage (CR2PA), L'archivage des mails ou Les utilisateurs face aux mails qui engagent l'entreprise. Livre blanc proposé par le groupe de travail "Archivage des mails" du CR2PA. Paris: CR2PA, 2009. https://blog.cr2pa.fr/wp-content/uploads/2013/02/CR2PA_Pub_Livre-blanc-Archivage-des-mails-2009_BR.pdf

[4] P. Ballet, J.M. Rietsch, Conserver les courriers électroniques? Ou comment résoudre la problématique de l'archivage des e-mails. Paris: FedISA, 2008. https://www.alain-bensoussan.com/wp-content/uploads/2542992.pdf

[5] Archives nationales, La gestion et l'archivage des courriels. Manuel pratique. Paris: August 2012 (2nd version). https://circgdquevilly.spip.ac-rouen.fr/IMG/pdf/vade-courriels.pdf
[6] C. J. Prom, "Preserving Email. DPC Technology Watch Report 111-01 December 2011". London: Digital Preservation Coalition, 2011
[7] S. Soyez, Directives pour la gestion et l'archivage numérique des e-mails. National Archives of Belgium – Algemeen Rijksarchief, October 2009 (version 1.1)
[8] Bibliothèque et Archives nationales du Québec, Orientations pour la gestion documentaire des courriels du gouvernement du Québec. 2009. https://numerique.banq.qc.ca/patrimoine/details/52327/2007081
[9] The National Archives, Guidelines on developing a policy for managing email. London: 2004
[10] F. Boudrez, Filing and archiving e-mail. Antwerp, 2006
[11] The National Archives of Australia, Managing Emails. https://www.naa.gov.au/information-management/types-information-and-systems/types-information/managing-email
[12] R. Ferrante, L. Schmitz Fuhrig, Digital Preservation: Using the Email Account XML Schema. Washington: Smithsonian Institution, [2007]
[13] The InterPARES 3 Project, TEAM Italy, Guidelines and Recommendations for E-Mail Records Management and Long-Term Preservation. 2011
[14] G. Knight, InSPECT Project Document - Significant Properties Testing Report: Electronic Mail. London: JISC/The National Archives/King's College London, 2009
[15] SIAF, SEDA standard Homepage. https://www.francearchives.fr/seda/
[16] C. J. Prom, "Preserving Email. DPC Technology Watch Report 19-01 May 2019". London: Digital Preservation Coalition, May 2019 (2nd Edition). https://www.dpconline.org/docs/technology-watch-reports/2159-twr19-01/file
[17] Council on Library and Information Resources, The Future of Email Archives. A Report from the Task Force on Technical Approaches for Email Archives. August 2018. https://www.clir.org/wp-content/uploads/sites/6/2018/08/CLIR-pub175.pdf
[18] National Archives and Records Administration, Email and Electronic Messages Management. 3 mars 2023 (last update). https://www.archives.gov/records-mgmt/email-mgmt
[19] National Archives and Records Administration. General Records Schedule 6.1: Email and Other Electronic Messages Managed under a Capstone Approach. General Records Schedule 6.1. January 2023. https://www.archives.gov/files/records-mgmt/grs/grs06-1.pdf
[20] B. Texier, "Archivage électronique: quand le Quai d'Orsay archive ses emails". Paris: Archimag, 21 August 2020 (last update: 28 September 2022). https://www.archimag.com/archives-patrimoine/2020/08/21/archivage-electronique-quai-orsay-archive-emails
[21] Mission des Archives auprès du Premier ministre, La gestion et la collecte d'archives électroniques. FAQ à destination des référents-archives des Services du Premier ministre. Paris: 2018. https://www.documentation-administrative.gouv.fr/adm-01858798/document
[22] C. Lefebvre, Le traitement des messages électroniques du fonds de la COP21. Conversion et extraction. Paris: Ministère de l'Europe et des Affaires étrangères, 14 March 2017
[23] E. Levasseur, M. Sin Blima-Barru, "Retour d'expérience sur la stratégie de préservation des Archives nationales". Paris: SIAF, 15 April 2022. https://siaf.hypotheses.org/1531
[24] Vitam, ReSIP Source code: https://github.com/ProgrammeVitam/sedatools/tree/master/resip
[25] Archives nationales, "Fiche – guide. Verser une messagerie aux Archives nationales". Paris: DINUC/DAD, 16 November 2021 (version 6.1)
[26] SIAF, Vade-mecum. Elaborer une charte "courriels" et l'inscrire dans une stratégie d'archivage. Paris: January 2015. http://www.piaf-archives.org/sites/default/files/vademecum_charte_courriel_V1.7.pdf
[27] DIAF, Note 2020/D/6412 "Orientations stratégiques à mettre en œuvre pour la conservation et l'archivage des messageries au sein des administrations de l'Etat". Paris: 18 May 2020
[28] Aristote, Pérennisation de l'information numérique – Cellule Formats Homepage. https://www.association-aristote.fr/cellule-format/
[29] V. Meynaud, L'archivage numérique des courriels au Conseil d'État. Réflexions fonctionnelles et techniques. Paris: 2021
[30] Ministère de la Culture et de la Communication, Ciel mon courrier. Charte sur les usages de la messagerie. Paris: [2014]
[31] Vitam, MailExtract Source code. https://github.com/ProgrammeVitam/mailextract
[32] Vitam, ReSIP. Paris: 4 March 2023. http://www.programmevitam.fr/ressources/DocCourante/autres/fonctionnel/VITAM_Manuel_ReSIP.pdf
[33] Ministères sociaux, Archifiltre-mails Source code. https://github.com/SocialGouv/archifiltre-mails
[34] Fabrique des ministères sociaux, Archifiltre-mails. 2023. https://archifiltre.fr/mails
[35] CADA, Avis n°20224983, 20226133, 20225331, 20226349, 20226355, 20226362. 2022
[36] RATOM, Review, Appraisal, and Triage of Mail Homepage. https://ratom.web.unc.edu/
[37] Stanford Libraries, ePADD Homepage. https://library.stanford.edu/projects/epadd
[38] E. Vasseur, B. Grailles, T. Aït El Mekki, "Improving the archiving and contextualization of electronic messaging in French". Ipres 2022, Online Innovation 2 OI2, 14 September 2022. https://www.dpconline.org/docs/miscellaneous/events/2022-events/2791-ipres-2022-proceedings/file p.374-377

Repository Staff Perspectives on the Benefits of Trustworthy Digital Repository Certification

Rebecca D. Frank
School of Information Sciences
University of Tennessee
USA
rfrank7@utk.edu
https://orcid.org/0000-0003-2064-5140

Abstract – This paper reports on the results from a qualitative study that asks whether and how staff members from TRAC certified repositories find value in the audit and certification process. While some interviewees found certification valuable, others argued that the costs outweighed the benefits or expressed ambivalence towards certification. Findings indicate that TRAC certification offered both internal and external benefits, such as improved documentation, accountability, transparency, communication, and standards, but there were concerns about high costs, implementation problems, and lack of objective evaluation criteria.

Keywords – Digital Preservation, Trustworthy Digital Repositories, ISO 16363, TRAC, Repository Assessment

Conference Topics – From Theory to Practice; Sustainability: Real and Imagined.

I. Introduction

Trustworthy Digital Repositories (TDRs) are organizations that are entrusted with the care and preservation of unique and valuable digital information. From research data, to government records, to cultural heritage information, these repositories ensure the longevity and accessibility of information on a global scale, e.g., [1].

Certification processes have been developed to ensure that the organizations entrusted with this valuable information are indeed able to carry out the work of long-term preservation. Audits carried out by external bodies administer and enforce these certification systems in order to provide assurance to stakeholders that the repositories are trustworthy.

The Trustworthy Digital Repositories: Audit and Certification (TRAC) process, which was strongly influenced by the ISO 16363 standard, is one such certification system [2], [3]. This process is a time-consuming and expensive undertaking for a digital repository, and can result in certification as trustworthy by a team of auditors managed by the Center for Research Libraries [3].

The earliest TRAC certification was issued in 2011, and the most recent in 2015, with an update issued in 2018 for one repository (i.e., CLOCKSS). The staff members of those repositories have therefore had time to reflect on the value proposition of TRAC certification. This paper, which is based on interviews with staff members from all six TRAC certified repositories, asks the following research questions:

- Do staff members from TRAC certified repositories find certification to be valuable?
- How do staff members from TRAC certified repositories characterize the value and/or benefits of TRAC certification?

My findings indicate that while many staff members from TRAC certified repositories find the audit and certification process to be valuable, and described concrete internal and external benefits, others described the process as more expensive than valuable, and some expressed ambivalence about TRAC certification.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

Despite the amount of time that has passed since these TRAC certifications, the ISO 16363 standard, which formed the foundation for the process, was approved in 2012 and was reviewed and confirmed in 2023. This means that current TDR certification processes that rely on, or are influenced by, ISO 16363 are using the same standard as the participants in this research.

II. Background

A. Trustworthy Digital Repositories & TRAC Certification

Trust is a central concept in digital preservation [4]–[6]. As early as 1996, members of the digital preservation community identified the need for a mechanism to ensure the trustworthiness of organizations entrusted with the care of unique and valuable digital information [6]. In the nearly 30 years since the Garrett and Waters report, several systems for the audit and certification of digital repositories have emerged, including TRAC, CoreTrustSeal, and nestor e.g., [2], [7], [8].

The TRAC system is based on the ISO 16363 standard, Audit and Certification of Trustworthy Digital Repositories [2]. This certification process is based on the Open Archival Information Systems (OAIS) Model [9], and repository certifications based on this standard have been administered by the Center for Research Libraries (CRL) and the Primary Trustworthy Digital Repository Authorisation Body (PTAB) [10]–[16].

The TDR certification process administered by CRL, TRAC, actively conducted audits from 2011 through 2015 and maintained the certifications awarded through those audits until at least 2018 [17]. The general process for TRAC certification involved repository staff members preparing documentation for review by a team of CRL auditors, followed by a site visit from a small group of auditors who would conduct interviews and inspections in order to assess the veracity of repository documentation [18]. A final determination would be made, and a report prepared for the repository with the findings from the audit team [11]–[16]. The TRAC certification system is the focus of this paper.

TDR certification, including TRAC as well as other systems such as CoreTrustSeal and nestor, is a phenomenon in need of further interrogation. In recent years, scholars such as Maemura, Moles, & Becker have argued that frameworks for repository assessment have not been sufficiently examined [19]. Scholarship about TDR certification has tended to focus on individual reports from organizations that engaged with certification in formal and informal ways e.g., [20]–[24]. Other publications have focused on the development and maintenance of the certification systems e.g., [25]–[29]. There is a need for research that takes a step back from development processes and individual implementations of certification systems to interrogate the value of TDR certification.

B. Benefits and/or Value of TDR Certification

Scholars who have examined the value of TDR certification such as Donaldson have focused on questions about the longevity of digital information in certified repositories, and how certified repositories present this information on their websites [30], [31]. Research has also developed a taxonomy that can be used to address questions about the societal impact of TDRs [32]. A 2018 iPres paper examined the benefits of certification in terms of the return on investment for a particular repository for both Data Seal of Approval and nestor certifications and found that stakeholder confidence, transparent documentation, and process improvement were the most important benefits for their organization [33].

Repositories that have achieved TRAC certification have written about the experience, presenting their certification as a positive development to repository stakeholders e.g., [34], [35], [24]. While much can be learned from this literature, it is unlikely that an organization would be critical of the system in a publication designed to promote their certification. Individuals and organizations involved in the creation of TDR systems have also published informative literature about those systems [25]–[28], [36]. The goal of this category of literature is often promotion of the certification systems, and therefore also has a particular point of view that is unlikely to be critical of TDR certification.

This paper builds upon the scholarship described above to ask whether and how staff members from TRAC certified repositories find value in the audit and certification process.

III. Research Methods

This paper is part of a larger research project whose goal is to understand risk for long-term preservation in the context of TRAC certification. The project involves interviews with standard developers, auditors, and staff members of TRAC certified repositories. In this paper, I report on the results of 21 interviews with repository staff members from repositories that have received TRAC certification. More information about the research methods, including data collection instruments and the code set used for analysis, is available Open Access at http://hdl.handle.net/2027.42/147539 [37].

A. Data Collection

At the time of data collection in 2016, there were six repositories with TRAC certification: Canadiana.org, Chronopolis, CLOCKSS, HathiTrust, Portico, and Scholars Portal. In-depth, semi-structured interviews were conducted with staff members from all six certified repositories, across three functional areas: repository administration/management, IT, and digital preservation. Previous research has demonstrated that the work of digital preservation involves collaboration across these areas [38].

The interviews, which lasted one to two hours, asked participants to discuss their experiences with the TRAC certification process, and to identify and discuss potential sources of risk for TDRs. Included in the interviews were questions about the cost, benefits, and value of TRAC certification. Audio recordings of the interviews were transcribed for analysis.

B. Data Analysis

Interview transcripts were coded using NVivo. For the first round of analysis, I used a combination of descriptive, analytic, and thematic codes. The code set consisted of codes addressing potential sources of risk, factors that influence the social construction of risk, the TRAC audit process, and attitudes about TDR certification. Working together with another coder to achieve an acceptable level of interrater reliability, we reached a Scott's pi of 0.711 for the subset of interviews with repository staff members [39], [40].

Secondary analysis was conducted by a single researcher, focusing on the topics of cost, benefit, and value of TRAC certification, and attitudes about TRAC certification.

IV. Findings

Findings from this research indicate that the value proposition of TRAC certification is still an open question. While some interviewees described TRAC certification as valuable, others argued that the costs outweighed the benefits. Some also expressed ambivalence about the value of certification. I have organized the findings into four sections based on my analysis: (A) internal benefits, (B) external benefits, (C) arguments that the cost outweighs the benefits, and (D) ambivalence about the benefits of TRAC certification.

A. Internal Benefits of TRAC Certification

TRAC certification was described as valuable for internal repository processes by 12 of the 21 interviewees included in this study.

When asked about the value of certification, interviewees explained that the audit process was valuable because: (1) it forced them to document their policies and practices; (2) the act of creating this documentation enabled them to develop a better understanding of their organization, and to establish a shared understanding of repository policies and practices across the entire organization; and (3) that the review of the documentation by external auditors created an added layer of accountability that ensured a higher quality of documentation than they would otherwise have produced.

The TRAC audit process requires that repositories provide extensive documentation of their policies and processes [3]. Repository Staff 03, 04, 07, 17, and 18 all explained that rather than providing existing documentation to the auditors, their organizations instead had to create current, up-to-date documentation for the purpose of the audit. For example:

> *"Going through the audit there were a lot of policies you have to have, and we sort of assumed we had them [but] we just didn't have them written down. Going through them we realized in a lot of cases we actually didn't have them."*
> *(Repository Staff 07)*

Similarly, Repository Staff 18 explained that the audit required his organization to formalize internal processes which were not previously documented:

> *"I think on the technical side, some of it was what we had. On the practice side, I think it was a good exercise 'cause it forced us to formalize some of these processes that we had done. But we had been doing it internally, but we hadn't actually said, 'Okay, well let's write down a step-by-step guide on how to do this.' And I think it was useful for us to internally self-organize the archive a little bit." (Repository Staff 18)*

The audit process created an incentive for the organization to create new documentation. For some this was a matter of articulating existing policies more clearly or updating older documentation. For others it meant that repositories had to create policies that did not previously exist. In some cases, the process of creating documentation revealed gaps that were previously unknown to repository staff members:

> *"[W]e used the same high-level classification of the threats, we certainly identified a lot of things at the operational level where we were not doing as good a job as we should have been. That was a big part of the value of the audit, was that it forced us to actually write down what the processes we were doing were supposed to be, and reviewing whether what was actually happening matched what was supposed to happen. And in many cases it didn't." (Repository Staff 13)*

Whether they had to create new documentation for the TRAC audit nor not, the act of gathering the required information into one coherent set of documents for auditors to review was described as beneficial. This activity created opportunities to share information across different functional areas within a repository, ensuring that the entire staff had a shared understanding of the mission, policies, and practices of the organization:

> *"[T]he audit process helped to make that a lot more concrete and to say here's what we're doing today. This is exactly what we're doing today. Here's the specifications, here's the metadata, here's the schematics. That's changed some over time as it should. That made it much more real for us … I think up to that, we'd been a little loosey-goosey. That we'll name file names however we want, right? We'll package them and name the packages however we want. That was the first step in my mind of making us much more of a professional organization. Where someone could come in from the outside and we could hand them a dump of stuff and they could actually figure out what we've got. That was a huge practical benefit for us." (Repository Staff 04)*

Interviewees described the review by external auditors as a benefit of certification. Specifically, they argued that there was an added layer of accountability that came with the auditors, in contrast to the limited accountability of a self-audit. For example, Repository Staff 07 said that the external auditors were helpful because the TRAC process did not leave room for the repository staff to skip over or take shortcuts for any of the requirements:

> *"I think that having a third party do the audit is much better because you can cheat a lot, inadvertently, when you're doing the self-audit. Just sort of say, "Oh yeah we've got that covered," without thinking it through. When you actually have to explain to a third party how you've got it covered, that's when you realize that maybe you don't." (Repository Staff 07)*

My findings indicate that the TRAC certification process led to internal benefits for certified repositories, including improved understanding of repository policies and practices, increased accountability because of the external auditors, and incentivized the creation of new documentation and formalization of internal processes.

B. External Benefits of TRAC Certification

Interviewees also discussed external benefits from TRAC certification. When discussing these benefits, repository staff members focused on what certification could help them communicate to outside parties, and the role that it allowed their repository to play in the digital preservation community. Nine of the 21 interviewees described TRAC certification as valuable specifically because it (1) improved the transparency of their organization; (2) facilitated communication with repository stakeholders; (3) gave them a competitive advantage in the recruitment of partners, sponsors, and/or funding; and/or (4) gave them an opportunity to be early adopters and establish standards for digital repositories.

Transparency is a central tenet of repository certification [41]. For example, Repository Staff 07 explained that the act of demonstrating trustworthiness by providing information about policies and practices improved his repository's transparency overall, and that the organization was more proactive about making this information publicly available after certification:

> *"I also think that there's ongoing value to having that kind of third party oversight in a formal way. But I also think that there is enough oversight now, and there's a lot more transparency on our part just in terms of us being proactive about publishing, and announcing these changes that we make over time, that I'm not as concerned about it." (Repository Staff 07)*

Another benefit of TRAC certification was the fact that certification was seen as communicating something important to repository stakeholders. For some, the goal of TRAC certification was to help stakeholders understand the capabilities of their repository. Both the certification itself, as well as the documentation that repository staff members prepared for the auditors were described as contributing to this benefit.

Repository Staff 07 described certification as a way to establish credibility with external stakeholders: "The reason for doing TRAC certification was to establish credibility in the area and we've done that." This interviewee went on to explain that they would only maintain certification if the organization could articulate a clear business reason for doing so: "because we're quite a small organization and because there's a significant investment of resources, we would certainly be open to doing it, it's just there would have to be an articulable business reason for doing it" (Repository Staff 07).

In addition to establishing credibility, TRAC certification was also described as something that provided reassurance to stakeholders, "No one has ever proactively asked for it, but when you mention, when I mention it, they shake their heads as though they are reassured in some vague, hard to define way" (Repository Staff 12).

TRAC certification was described as a way to gain a competitive advantage by some interviewees. For repositories with active dues-paying members, for example, certification was viewed as a way to differentiate their organization from others and demonstrate their value. Repository Staff 11 said that the certification helped to recruit members: "[I]t has been useful for us to be able to say that we are certified. It's been useful to be able to say that to libraries and to [partners]. In terms of really practical areas, one of the things we've found is that sort of unexpectedly it brought some new [partners] to us."

Other repository staff members framed this benefit not as a way to recruit or maintain partners specifically, but rather as a necessary credential to maintain an overall competitive advantage. Repository Staff 13 was confident that his repository would lose business if they did not become TRAC certified: "It was a competitive threat … Without it [repository] would have lost business." (Repository Staff 13)

The repositories included in this study were early adopters of repository certification. This was explicitly described as a benefit. Interviewees

explained that it was important for their organizations to contribute to the establishment of standards in digital preservation by stepping up to go through this new audit process:

> *"It also seemed, to me and the team I think, important for us being part of the larger preservation community. I believed, and I believe now, that preservation of electronic materials is a really important effort, and a relatively new one, still today. Just going through the TRAC audit and taking, once I think, the risk of being [an early] enterprise to go through a TRAC audit, so scary, but potentially just so important for the community." (Repository Staff 08)*

These findings demonstrate the ways in which interviewees described external benefits of TRAC certification that focused on what the certification could communicate to external stakeholders, and the role it allowed them to play in the digital preservation community broadly.

C. The Cost of TRAC Certification Outweighs the Benefits

In contrast, six of the 21 interviewees argued that TRAC certification was not valuable for their repository because: (1) the high cost of certification outweighed the benefits; (2) TRAC is not well-known enough to be meaningful; (3) they found problems with the way that certification was implemented.

TRAC certification was described by all the interviewees in this research as very expensive, both in terms of money as well as the time that staff members had to spend preparing documentation for the auditors. Some were skeptical about whether these costs outweighed any benefits that they received from certification: "I doubt that the benefits outweigh the costs. I'm sorry to say that. It is not clear to me that the benefits are worth the costs" (Repository Staff 08).

Similarly, Repository Staff 13 said that the TRAC audit process was both costly and disruptive for his repository: "I think there are really big issues about how expensive and disruptive the process is, relative to the benefits that you gain from it. Because there clearly are benefits, but the costs and the disruption are very large" (Repository Staff 13). This interviewee went on to explain that he believed that his organization could have found less costly ways to get the benefits from certification, but that they felt that certification was necessary for financial reasons, "we were under significant competitive pressure. If it hadn't been for that, we could have got most of these benefits at much lower cost by a more gradual approach, rather than going all the way to TRAC in one go" (Repository Staff 13).

Repository Staff 04 and 11 both felt that the costs of TRAC certification would be barrier for future adoption. Repository Staff 04 argued that cost would need to be lowered substantially for certification to be viable, because the process was prohibitively expensive for his repository. And Repository Staff 11 said that it would take a significant amount of external pressure from stakeholders to go through another audit for recertification: "Honestly, it was such a pain in the butt I am not anxious to do it again. If we started getting pressure from CRL or our libraries or our publishers, then I suspect we would. Without that, my instinct is to coast, actually. It was so much work that, man, we'd have to have a good reason to do it again."

Repository Staff 13 explained that he would steer any organizations affiliated with his repository away from TRAC certification because of the costs: "There's no need for any of the other [affiliated] archives at the moment to get certified, and if there was, I wouldn't recommend that they get 16363, because of the resource implications of trying to do it."

While the opportunity to be an early adopter was described as a benefit of certification by repository staff members in this research, the relative newness of certification was also seen as a drawback. For Repository Staff 20, a major drawback of certification was that it was relatively unknown and so would not necessarily communicate effectively about his repository's trustworthiness to others because they would not know what it meant to be TRAC certified: "nobody in [country] had been certified as a Trusted Digital Repository before. So, it was more like not even the process itself but the fact that it was kind of an unknown thing" (Repository Staff 20).

There were several issues raised about the way that TRAC certification was implemented.

Interviewees argued that the requirements of TRAC certification were not stringent enough, and that the OAIS model on which TRAC certification was based failed to address the realities of managing a digital repository: "for us the TRAC certification was particularly tricky because TRAC is totally based on OAIS, which totally does not understand a number of aspects of running real world repositories" (Repository Staff 13).

Repository Staff 02 explained that there are no minimum thresholds in TRAC and that repositories could become certified with lots of caveats for sub-optimal policies/practices. Indeed, other research has found that repositories were able to become TRAC certified without fully meeting the requirements outlined in the checklist [18], [42].

Repository Staff 08 expressed dissatisfaction with the auditors. She argued that they were less knowledgeable about digital preservation than the staff of her own repository, which made her doubtful about whether the certification itself held meaning:

> *"I guess one of my take-aways from the TRAC audit at [repository], and this is my own personal opinion - ... Take it for what it's worth, when I weigh the level of expertise of the operational team at [repository] against the amount of time and effort put into the documentation used by auditors who in my opinion, please forgive me, were significantly less expert, it made me concerned about the value of the outcome." (Repository Staff 08)*

Six out of 21 interviewees in this study discussed the costs or drawbacks of TRAC certification. They argued that the costs of TRAC certification outweighed the benefits, that TRAC was not well-known enough to be meaningful, and described problems with the way certification was implemented. All interviewees described TRAC certification as very expensive, both in terms of money and time spent preparing documentation for auditors. Some interviewees felt that the costs of certification would be a barrier for future adoption.

D. Ambivalence about the Value of TRAC Certification

In contrast to Sections A, B, and C above in which interviewees argued for or against the value of certification, some interviewees were ambivalent about TRAC. In particular, interviewees were skeptical about the usefulness of the audit outcomes. Interviewees argued that the lack of objective evaluation criteria meant that audit scores were not meaningful and therefore could not be used, for example, to compare their organization against others.

Repository Staff 16 explained that rather than evaluating repositories against an objective set of criteria, the process was designed to assess whether each individual repository was in fact operating in accordance with their own policies: "[T]hey certify that you do what you say you do. They don't certify that you do something good. Which is a little bit of a vague. So how good you are is what you decide to document and what you decide the processes to be" (Repository Staff 16).

Repository Staff 11 also discussed the flexibility of the TRAC requirements. This interviewee explained that the flexibility was frustrating because it meant that the scores issued for each repository were essentially meaningless and could not be compared against one another:

> *"That's one of the interesting things about TRAC, right, is that level of flexibility. It's also sort of one of the frustrating things about it, too. Because, you know, grades aren't equal. One institution's score doesn't mean they're providing the same level of preservation as another institution's score, because you're evaluating the institution against what the institution said it would do, not against some yardstick." (Repository Staff 11)*

For these interviewees, the value of TRAC certification was an open question, because the scores themselves were viewed as lacking meaningful information about how each repository compared with others. This is particularly interesting in light of the findings from section B above, in which interviewees argued that a key benefit of TRAC certification was that the results facilitated communication with repository stakeholders and

conferred a competitive advantage on certified repositories.

Ambivalence about TRAC focused on what the certification could *not* communicate externally about certified repositories. Staff members from certified repositories believed that it should differentiate their organizations from others by demonstrating their trustworthiness and compliance with best practices. They were frustrated to learn that the results of the process could not be used to make direct comparisons, and that repositories with vastly different practices could receive similar scores.

The flexibility of the TRAC requirements was frustrating for interviewees, as they believed that this meant that the scores issued for each repository were essentially meaningless and could not be compared against one another.

V. Discussion & Conclusion

This study highlights the complex and varied perspectives on the value of TRAC certification for digital repositories and the need to continue to examine how certified repositories view the value and benefits of the process over time. My findings indicate that staff members of TRAC certified repositories understood certification to have both internal and external benefits for repositories, including improved transparency, communication with stakeholders, and a competitive advantage in recruitment of partners, sponsors, and funding. However, some interviewees argued that the high cost of certification outweighed the benefits, that TRAC was not well-known enough to be meaningful, and that there were problems with the way certification was implemented. Additionally, some interviewees expressed ambivalence about the value of TRAC certification, arguing that the lack of objective evaluation criteria meant that the audit scores were not meaningful.

This aligns with findings from my previous research in which I found that the highly flexible certification criteria, which are intended to allow the system to be applicable across a broad array of repository types, have been used by repositories to justify sub-optimal preservation practices e.g., [18], [42]. In this paper I argue that this flexibility, which I have characterized elsewhere as a potential source of risk for both digital repositories and the long-term preservation of the digital information they contain, also detracts from the value of certification for some stakeholders.

This study complements previous research about the value of TDR certification. For example, Donaldson has carried out research which seeks to understand whether repositories with TDR certification have better long-term outcomes, in order to understand the impact of certification [32]. Notably, my findings show that despite the benefits listed here, participants did not say that the information in their repositories was more secure or better preserved after completing a TRAC audit. Also absent were arguments that their repositories were more trustworthy or better able to preserve information long term as a result of going through the TRAC audit process. Rather, the benefits centered on aligning the expectations of internal and external stakeholders, and of improving transparency and communication in order to remain competitive.

As discussed in Section II B above, much of what is known about TDR certification has been produced by those involved in the process in some way – developers of certification systems, and repositories that have achieved certification. This paper provides a new perspective, investigating the value of TRAC certification through empirical research. Even so, participants in this study may still have been motivated by a desire to promote the certification system. Achieving TRAC certification was a costly endeavor and phenomena such as escalation of commitment and/or sunk cost bias may have been present in this study [43], [44].

Future research, which considers both the repository outcomes as well as the attitudes and beliefs of repository staff members has the potential to produce a more complete picture of the value of this relatively new phenomenon. Additionally, as more time passes, repository stakeholders may be willing and/or able to reflect on their experiences with TDR certification in different ways.

Finally, TRAC is one of several TDR certification systems that are active today. While some of the criticism about TRAC certification focused on the requirements themselves, much centered on the particular implementation of TRAC certification as administered by CRL. More recent audits have been conducted by a different organization (i.e., PTAB),

and future research should investigate this new implementation of the ISO 16363 standard [10].

VI. Acknowledgments

I would like to thank Megh Marathe and Carl Haynes for their assistance with data analysis. Thank you also to Elizabeth Yakel, Ph.D. for support and thoughtful feedback across the lifecycle of the project. This research was funded in part by a University of Michigan Rackham Graduate Student Research Grant.

1. References

[1] CoreTrustSeal, "Core Certified Repositories," *CoreTrustSeal*, 2021. https://www.coretrustseal.org/why-certification/certified-repositories/ (accessed Oct. 17, 2021).

[2] Consultative Committee for Space Data Systems, "Audit and Certification of Trustworthy Digital Repositories," Consultative Committee for Space Data Systems, Washington, D.C., Standard ISO 16363:2012 (CCSDS 652-R-1), 2012. Accessed: Aug. 05, 2013. [Online]. Available: http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510

[3] Center for Research Libraries, "TRAC Metrics," *CRL: Center for Research Libraries Global Resources Network*. https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/trac (accessed Mar. 07, 2023).

[4] P. E. Hart and Z. Liu, "Trust in the Preservation of Digital Information," *Commun ACM*, vol. 46, no. 6, pp. 93–97, Jun. 2003, doi: 10.1145/777313.777319.

[5] G. Bak, "Trusted by Whom? TDRs, Standards Culture and the Nature of Trust," *Arch. Sci.*, vol. 16, no. 4, pp. 373–402, Dec. 2016, doi: 10.1007/s10502-015-9257-1.

[6] J. Garrett and D. J. Waters, "Preserving Digital Information: Report of the Task Force on Archiving of Digital Information," The Commission on Preservation and Access & Research Libraries Group, Washington, D.C., 9781887334501 1887334505, 1996. [Online]. Available: https://www.clir.org/wp-content/uploads/sites/6/pub63watersgarrett.pdf

[7] CoreTrustSeal Standards and Certification Board, "CoreTrustSeal Requirements 2023-2025," Sep. 2022, doi: 10.5281/zenodo.7051012.

[8] nestor Working Group Trusted Repositories - Certification, "nestor Criteria: Catalogue of Criteria for Trusted Digital Repositories, Version 2," Deutsche Nationalbibliothek, Frankfurt am Main, Nov. 2009. [Online]. Available: http://nbn-resolving.de/urn:nbn:de:0008-2010030806

[9] Consultative Committee for Space Data Systems, "Reference Model for an Open Archival Information System (OAIS)," Consultative Committee for Space Data Systems, Washington, D.C., Magenta Book CCSDS 650.0-M-2, 2012. Accessed: Jul. 19, 2022. [Online]. Available: https://public.ccsds.org/Pubs/650x0m2.pdf

[10] PTAB - Primary Trustworthy Digital Repository Authorisation Body Ltd., "Certified clients," *PTAB - Primary Trustworthy Digital Repository Authorisation Body Ltd*, 2021. http://www.iso16363.org/iso-certification/certified-clients/ (accessed May 26, 2021).

[11] Center for Research Libraries, "CRL Certification Report on CLOCKSS Audit Findings," Center for Research Libraries, 2014. Accessed: Aug. 11, 2014. [Online]. Available: http://www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories/clockss-report

[12] Center for Research Libraries, "CRL Certification Report on Portico Audit Findings," Center for Research Libraries, Chicago, IL, 2010. [Online]. Available: https://www.crl.edu/sites/default/files/reports/CRL%20Report%20on%20Portico%20Audit%202010.pdf

[13] Center for Research Libraries, "CRL Certification Report on Chronopolis Audit Findings," Center for Research Libraries, Chicago, IL, 2012. [Online]. Available: https://www.crl.edu/sites/default/files/reports/Chron_Report_2012_final_0.pdf

[14] Center for Research Libraries, "CRL Certification Report on the Canadiana.org Digital Repository," Center for Research Libraries, Chicago, IL, 2015. [Online]. Available:

https://www.crl.edu/sites/default/files/reports/CANADIANA_AUDIT%20REPORT_2015.pdf

[15] Center for Research Libraries, "CRL Certification Report on the HathiTrust Digital Repository," Center for Research Libraries, Chicago, IL, 2011. [Online]. Available: https://www.crl.edu/sites/default/files/reports/CRL%20HathiTrust%202011.pdf

[16] Center for Research Libraries, "CRL Certification Report on Scholars Portal Audit Findings," Center for Research Libraries, Chicago, IL, 2013. Accessed: May 01, 2019. [Online]. Available: http://www.crl.edu/sites/default/files/attachments/pages/ScholarsPortal_Report_2013_%C6%92.pdf

[17] Center for Research Libraries, "2018 Updated Certification Report on CLOCKSS," Center for Research Libraries, Chicago, IL, 2018. [Online]. Available: https://www.crl.edu/sites/default/files/reports/CLOCKSS_Report_2018_0.pdf

[18] R. D. Frank, "Risk in Trustworthy Digital Repository Audit and Certification," *Arch. Sci.*, vol. 22, no. 1, pp. 43–73, Mar. 2022, doi: 10.1007/s10502-021-09366-z.

[19] E. Maemura, N. Moles, and C. Becker, "Organizational assessment frameworks for digital preservation: A literature review and mapping," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 7, pp. 1619–1637, 2017, doi: 10.1002/asi.23807.

[20] B. Houghton, "Trustworthiness: Self-assessment of an Institutional Repository against ISO 16363-2012," *-Lib Mag.*, vol. 21, no. 3/4, Mar. 2015, doi: 10.1045/march2015-houghton.

[21] A. Krahmer and M. E. Phillips, "Communicating Organizational Commitment to Long-Term Sustainability through a Trusted Digital Repository Self-Audit," presented at the IFLA WLIC 2016: Connections. Collaboration. Community, Columbus, OH, 2016. Accessed: Mar. 07, 2023. [Online]. Available: https://library.ifla.org/id/eprint/1505/

[22] A. Krahmer, P. Andrews, H. Tarver, M. E. Phillips, and D. Alemneh, "Documenting Institutional Knowledge Through TRAC Self-Audit: A Case Study," in *Knowledge Discovery and Data Design Innovation*, Dallas, Texas, USA: WORLD SCIENTIFIC, Dec. 2017, pp. 335–348. doi: 10.1142/9789813234482_0018.

[23] A. M. Medina-Smith, "A Self-Audit of the NIST Public Data Repository Using the CoreTrustSeal Trustworthy Data Repositories Requirements," *NIST InteragencyInternal Rep. NISTIR - 8341*, Apr. 2021, doi: https://doi.org/10.6028/NIST.IR.8341.

[24] A. Kirchhoff, E. Fenton, S. Orphan, and S. Morrissey, "Becoming a Certified Trustworthy Digital Repository: The Portico Experience," in *Proceedings of the 7th International Conference on Preservation of Digital Objects*, Vienna, Austria, 2010, pp. 87–94. [Online]. Available: https://phaidra.univie.ac.at/o:185497

[25] I. Dillo and L. De Leeuw, "CoreTrustSeal," *Mitteilungen Ver. Österr. Bibl. Bibl.*, vol. 71, no. 1, pp. 162–170, Jul. 2018, doi: 10.31263/voebm.v71i1.1981.

[26] S. Dobratz and A. Schoger, "Trustworthy Digital Long-Term Repositories: The nestor Approach in the Context of International Developments," in *Research and Advanced Technology for Digital Libraries*, L. Kovács, N. Fuhr, and C. Meghini, Eds., in Lecture Notes in Computer Science, no. 4675. Springer Berlin Heidelberg, 2007, pp. 210–222. Accessed: Aug. 04, 2014. [Online]. Available: http://link.springer.com/chapter/10.1007/978-3-540-74851-9_18

[27] D. Giaretta, "OAIS Model and Certification of Trusted Digital Repositories," Fondazione Rinascimento Digitale, 2012. Accessed: Aug. 11, 2014. [Online]. Available: http://93.63.166.138:8080/dspace/handle/2012/117

[28] H. L'Hours, M. Kleemola, and L. De Leeuw, "CoreTrustSeal: From Academic Collaboration to Sustainable Services," *IASSIST Q.*, vol. 43, no. 1, pp. 1–17, May 2019, doi: 10.29173/iq936.

[29] E. Zierau *et al.*, "OAIS Version 3 Draft Updates: The 16th International Conference on Preservation of Digital Objects," in *Proceedings of The 16th International Conference on Preservation of Digital Objects*, Amsterdam, Netherlands, Sep. 2019, pp. 254–259.

[30] D. R. Donaldson, I. Dillo, R. Downs, and S. Ramdeen, "The Perceived Value of Acquiring

Data Seals of Approval," 2017, doi: https://doi.org/10.2218/ijdc.v12i1.481.

[31] D. R. Donaldson, "Certification Information on Trustworthy Digital Repository Websites: A Content Analysis," *PLoS ONE*, vol. 15, no. 12, p. e0242525, Dec. 2020, doi: 10.1371/journal.pone.0242525.

[32] D. R. Donaldson and S. V. Russell, "Towards a Taxonomy of Trustworthy Digital Repository Impacts," *Proc. Assoc. Inf. Sci. Technol.*, vol. 58, no. 1, pp. 430–434, 2021, doi: 10.1002/pra2.473.

[33] M. Lindlar and F. Schwab, "All that work … for what? Return on investment for trustworthy archive certification processes – a case study.," in *Proceedings of the 15th International Conference of Digital Preservation*, Boston, MA: Open Science Framework, 2019. doi: 10.17605/OSF.IO/8A3SC.

[34] CLOCKSS, "CLOCKSS Archive Certified as Trusted Digital Repository; Garners top score in Technologies…," *CLOCKSS News*, Jul. 28, 2014. https://www.clockss.org/clockss/News (accessed Mar. 30, 2016).

[35] D. Free, "HathiTrust Certified Trustworthy Repository," *Coll. Res. Libr. News*, vol. 72, no. 5, p. 254, 2011.

[36] S. Dobratz, A. Schoger, and S. Strathmann, "The nestor Catalogue of Criteria for Trusted Digital Repository Evaluation and Certification," *J. Digit. Inf.*, vol. 8, no. 2, Sep. 2007, Accessed: Apr. 06, 2013. [Online]. Available: http://journals.tdl.org/jodi/index.php/jodi/article/view/199/180

[37] R. D. Frank, "The Social Construction of Risk in Trustworthy Digital Repository Audit and Certification," Dissertation, University of Michigan, Ann Arbor, MI, 2018. Accessed: Oct. 01, 2019. [Online]. Available: https://deepblue.lib.umich.edu/handle/2027.42/147539?show=full

[38] R. D. Frank and E. Yakel, "Disaster Planning for Digital Repositories," in *Proceedings of the American Society for Information Science and Technology*, Montreal, QC, Canada, 2013, pp. 1–10. doi: 10.1002/meet.14505001058.

[39] R. T. Craig, "Generalization of Scott's Index of Intercoder Agreement," *Public Opin. Q.*, vol. 45, no. 2, pp. 260–264, 1981, doi: 10.1086/268657.

[40] W. A. Scott, "Reliability of Content Analysis: The Case of Nominal Scale Coding," *Public Opin. Q.*, vol. 19, no. 3, p. 321, 1955, doi: 10.1086/266577.

[41] B. F. Reilly, Jr. and M. E. Waltz, "Trustworthy Data Repositories: The Value and Benefits of Auditing and Certification," in *Research Data Management : Practical Strategies for Information Professionals*, J. M. Ray, Ed., Ashland, OH: Purdue University Press, 2013, pp. 109–126.

[42] R. D. Frank and L. Rothfritz, "Designated Community: uncertainty and risk," *J. Doc.*, vol. 79, no. 4, pp. 880–897, May 2023, doi: 10.1108/JD-07-2022-0161.

[43] K. A. Dijkstra and Y. Hong, "The feeling of throwing good money after bad: The role of affective reaction in the sunk-cost fallacy," *PLOS ONE*, vol. 14, no. 1, p. e0209900, Jan. 2019, doi: 10.1371/journal.pone.0209900.

[44] M. Kajtazi, H. Cavusoglu, I. Benbasat, and D. Haftor, "Escalation of commitment as an antecedent to noncompliance with information security policy," *Inf. Comput. Secur.*, vol. 26, no. 2, pp. 171–193, Jun. 2018, doi: 10.1108/ICS-09-2017-0066.

A Question of Character

How do we automatically recharacterize data at cloud scales?

Jack O'Sullivan
Preservica Ltd
UK
jack.osullivan@preservica.com
https://orcid.org/0000-0002-0306-761X

David Clipsham
Preservica Ltd
UK
david.clipsham@preservica.com
https://orcid.org/0009-0006-2611-8877

Divyesh Soni
Preservica Ltd
UK
divyesh.soni@preservica.com

Richard Smith
Preservica Ltd
UK
richard.smith@preservica.com

Jonathan Tilbury
Preservica Ltd
UK
jonathan.tilbury@preservica.com

Abstract – Many preservation actions that we undertake on digital content are driven by the format of the content in question. Format information is often determined at the point of ingest and is not regularly updated as our knowledge of file formats improves over time. Periodically re-characterizing all content in a repository would ensure that we get more accurate identifications over time, but a more sustainable approach would be to only re-characterize content that was actually likely to have changed. Preservica's new Automated Active Digital Preservation feature seeks to do exactly this, but even when considering only subsets of the data in our cloud systems, we are faced with significant challenges of scale. In this paper, we describe those challenges, the approach we have taken to implement the feature, and the testing we have performed to verify the viability of this approach.

Keywords – Scalability, Automation, Characterization, Preservation Actions

Conference Topics – From Theory to Practice; Sustainability: Real and Imagined.

I. Introduction

Characterization is one of the fundamental bases of Digital Preservation. It is the process of identifying the types of digital material we are preserving, and extracting the relevant technical characteristics and significant properties of that material [1]. This understanding of our content drives many digital preservation processes and policies; it might inform how and where we store the content, what normalizations, if any, we perform, what access copies we need to generate, and how we display content to end users. Its importance is such that it is an assumed standard part of our digital preservation processes, with at least the identification part of it even being part of the "Parsimonious Preservation" workflow [2].

Characterization is often treated as part of the ingest process, or preparation for the ingest process [3], and it is true that performing characterization up-front has benefits. Until we know what our digital material is, we can't apply format based policies, or take format based preservation actions such as normalization or the creation of access copies. However, our collective understanding and knowledge of file formats changes over time, as do the tools available to identify and validate content, and to perform extraction of technical properties. If all we have is the knowledge of how our content was identified at the point of ingest, and the characteristics we could measure with the tools then available, then our decision making about all subsequent preservation actions may be flawed.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

Ideally, our content should be characterized with the latest file format knowledge and most up to date tools at all times.

If re-characterization is a process that must be manually undertaken, this places a burden on the user/s of the system to ensure that this happens. These users are often archivists and collection managers rather than digital preservation experts, and as such are not always the people best placed to determine what needs to be re-characterized and what does not.

An alternative approach would be to automate re-characterization on a periodic basis, in the way that we might perform fixity checks, in order to ensure that our information up to date. However, this potentially requires a lot of compute time, and will, more often than not, result in no changes needing to be made.

Preservica has developed a feature that ensures that the preservation system itself can automatically respond to recommendations made by digital preservation experts to ensure that the correct subset of repository content is re-characterized as appropriate. This removes that burden from non-expert users of our systems, and means we only run processes on potentially affected content.

In this paper, we will discuss how even this approach results in challenges of scale when applied to production systems. In section II we will discuss what these challenges are. In sections III and IV we will discuss our approach and what steps we took to verify that it would work at the scales required and in section V we will discuss how well this matched the performance we saw when taking this feature into production.

II. What Do We Mean By Scale

A. Scale of the Format Problem

A blog post in 2018 [4], investigated the specific case of how PDF identification within PRONOM and DROID had evolved and demonstrated that the identification outcome of a corpus of PDF files changed over time. This is a natural consequence of the fact that PRONOM's data changes over time, usually for the better, as PRONOM's global community of contributors feedback their expertise into the dataset.

This was explored further in a poster for iPres 2019 [5] which additionally examined historical changes to the GIF, TIFF, and JPEG PRONOM-based identification.

However, PRONOM contains details of over 2250 file formats as of March 2023, so it is necessary to evaluate changes across the entire dataset to get a complete understanding of the impact of these changes.

Carrying on from the Lightning Talk last year [6], we investigated changes in PRONOM going back to the very earliest versions, with the PRONOM v10 update in 2006 chosen as a starting point as this was the first release where every single format entry had a persistent 'PRONOM Unique Identifier' (PUID) assigned.

Of an initial assessment of 1,089 unique file formats represented across the Preservica Cloud estate as of March 2022, we found that 489 format definitions (approximately 45% of those assessed) have changed at least once in such a manner that they warrant a re-identification event.

All of these recommendations have been made publicly available and as new recommendations are made these will continue to be published for the benefit of all.

Format definitions change in PRONOM for a few reasons:

Name or version updates: These are often relatively trivial, so a format name might be updated to correct a misspelling or to match official branding. A format version might be adjusted to cover multiple software releases or adjusted to a default 'generic' entry that is used in the event of a format being unable to be identified as an exact, specific version. There can be more impactful changes, however.

In the case of the database preservation file format, the Software-Independent Archiving of Relational Databases format, or SIARD, when the format was originally added to PRONOM in 2009 the entry was given the version number 2, although version 2 of the format wasn't formalized as a standard until 2015. In 2014, on the advice of the Swiss Federal Archives who created the original file format, the original entry in PRONOM was adjusted to version 1.0. Subsequently in 2016 SIARD version 2.0 was added to PRONOM. As such two separate

PRONOM entries have been called 'SIARD 2.0' at separate times, therefore file instances that were most recently identified before the 2014 correction will need to be re-identified to ensure they have the correct identification and to avoid confusion and ensure proper management.

In a separate case, the image file format '3D Studio,' introduced in one of the earliest versions of PRONOM before version 10, had its name changed to 'Paint Shop Pro Image' for reasons unknown around 2012. This was likely a mistake, as it was changed back to '3D Studio' in 2015 but this means that any file instances identified as such during this time period will need to be re-identified.

Up to the version 109 PRONOM update in November 2022, 301 updates to format name and/or version number have taken place.

Deprecations: Once a PRONOM entry has been created, it is intended to persist, so entries are not permanently deleted for any reason, however sometimes an entry may no longer be suitable for use, at which point it is deemed 'deprecated' and disassociated from identification mechanisms such as extension or file format signatures. Particularly in the early days of PRONOM there were several entries added that really related to specific software versions rather than file format versions and subsequent research deemed many of these unnecessary and with the potential to cause unintentional and unwanted identification clashes.

In the case of the Tagged Image File Format, or TIFF, PRONOM originally had distinct entries for versions 3, 4, 5, and 6, however each entry shared a single identification signature, meaning a file format identification tool would identify a file instance as each of these four formats, which could cause confusion or uncertainty, however it wasn't clear how to distinguish between these format versions reliably. A decision was made to deprecate these four entries and create a single general one. As such any file instances that were identified before these deprecations were made, should be re-identified to ensure they get the current correct identification outcome.

As of PRONOM's version 109 update, 68 file format entries have been deprecated.

Changes to format priorities: Further significant sources of change within PRONOM are 'priority relationships.' Many file formats are based on other file formats and some formats share certain characteristics of others. In these cases, it may be the case that these shared characteristics, where used for file format identification, will clash and would result in a file format identification tool matching against each format rather than a specific one. This situation is handled through setting a 'priority relationship,' where the more specific format is given priority over the more general one.

A new priority relationship being introduced will usually necessitate some form of re-identification as the previously general format identification outcome may now result in a more specific outcome if reassessed. A common case is where camera image formats, such as the Nikon NEF, the Pentax PEF, and similar file formats which are often based upon the TIFF file format, are introduced. Since these would have previously been identified as TIFF, it follows that any previously identified TIFF files should be re-identified as these may now get a more specific identification outcome. This is an instance that would need to be handled with care however, as many digital preservation repositories will store many millions of TIFF files.

In a separate case, when the Video Object Format (VOB) was introduced to PRONOM in 2012, it was given a *lower priority* than the MPEG Program Stream video formats from which it was derived. This was a mistake, as VOB is the more specific format so it should have been given a higher priority. This mistake was corrected in 2014 but means that any file instances that were identified as MPEG-1 or MPEG-2 Program Stream during this time need to be re-identified as they may have instead been VOB files.

As of the version 109 PRONOM update there are 1,054 priority relationships in-place, with 191 formats set as 'lower priority' than one or more other formats.

Changes to identification signatures: The final major trigger for file format re-identification will be where file format identification signatures are changed.

This usually happens where a previous signature has been found to be a little loose in order to tighten the signature, however it can sometimes be the opposite, where a previous signature has been a

little too strict. This could also be correcting a prior mistake.

A signature update will not necessarily require a new re-identification as in many cases optimizing a signature will not adversely affect a prior identification outcome, but mistakes will usually necessitate them.

In a recent instance, an attempt to slightly loosen up the signature for Encapsulated PostScript (EPS) version 2.0 went awry – the intention was to replace three specific bytes with wildcard bytes (bytes that can have any value) to allow for a little variance that had been observed in some file instances. Mistakenly the sequence was replaced with two wildcard bytes rather than three, which meant that affected files would then erroneously identify as standard PostScript rather than Encapsulated PostScript. This issue was quickly rectified within two months, but once again, any file instances that were identified as PostScript during this time will need to be re-identified.

From version 10 to the version 109 PRONOM update, 594 signature sequences have been altered.

B. Scale of the Content Problem

Preservica has been running commercial, cloud-based digital preservation systems for over a decade; starting with a single, multi-tenant system in the US, we now operate tens of systems across multiple regions of the world. Some of these are "private cloud" systems, hosting services and data for a single organization, others are multi-tenant, with tens, hundreds and even thousands of organizations sharing resources. We have customers who have been using these systems continuously for the entire lifetime of the service, meaning that we have production data that was ingested over ten years ago.

As of October 2022, we have over 116 million digital objects stored across our cloud estate. Our largest individual tenancies each have over 10 million assets stored.

Of these files there are approximately 1,350 file formats represented across the estate. The top ten most common file formats present make up over 90 million assets, approximately 77% of files stored. The most common types of file format present are images, documents (including PDF), and email.

We have over 32 million TIFF files stored, and a similar number of the various JPEG file format variants. There are over 20 million PDFs, including over 2 million PDF/A files. There are approximately 4.5 million emails.

However, the long tail is very real and very long. 664 file formats have 100 or fewer assets stored. 1,056 file formats have fewer than 1,000 assets. The 1,000 least populous file formats make up just under 110,000 files stored, less than 1% of the total, and although 1% seems like a very small number, 110,000 is more files than many of our individual tenants have in total.

The diversity of file formats present truly reflects the diversity of our user-base. Among these file formats we see rare and interesting eBook formats such as Broad Band LRF, or the Rocket Book eBook format. We see many different variants of Flash, which was once extremely common but due to security issues is no longer supported by most mainstream content platforms. We see ancient image formats such as PCX and TGA, but also extremely modern ones such as HEIF and JPEG XL.

Some proportion of this content will have been tentatively identified. This means that it didn't match any byte sequences for any file formats, and was assigned an identification on the sole basis of the file extension. Whilst we know this must be true for some content (e.g. the plain text file format x-fmt/111 has no byte sequences to match), the raw format data we have analyzed does not tell us this for other formats where byte sequences do exist. Some of the changes made to PRONOM in the time since any such content was ingested might mean that today we would be able to provide a firmer identification on the basis of matching byte sequences.

For example, the OS/2 Presentation Manager Metafile file format was originally added to PRONOM in 2005, and was associated with the .met extension so any file instances with that extension will have received a tentative identification outcome. In the v108 PRONOM update in 2022, a new identification signature for this file format was created, meaning we can now re-identify these file instances and either definitively and positively identify them as OS/2 Metafiles, or for those that are not OS/2 Metafiles, focus file format identification research efforts to further improve PRONOM.

We also have approximately 700,000 (approximately 6% of the total) 'unidentified' file formats stored, that is files for which we were unable to positively assert the file format identity *at the point of ingest*. Since the time period for these ingests stretches many years and PRONOM coverage is continually improving, the real current number is likely to be lower, but this can only be measured through re-identification.

These counts are only looking at the cloud services that Preservica actively manages. We have a number of "on-premise" customers who themselves manage similar sized repositories. Our on-premise offering pre-dates our cloud offering by around a decade, and so some of these customers have content ingested over even longer timescales.

III. Process

The approach we have taken to this problem of re-characterization at such scales is to separate responsibility for determining what content needs to be re-characterized from responsibility for actually running the process. Further, we have removed both responsibilities from the typical non-expert users of Preservica.

A. Identifying Changes

Preservica now allows a Digital Preservation expert to produce "Recommended Processes" [7], which describe the type of process to run and filters to describe the subset of content to run against. These filters include:

- lists of file formats to specify that only content matching one of the formats should be processed;
- event/date ranges to specify that only content ingested or last characterized between certain dates should be processed;
- whether unidentified, or tentatively identified content should be processed.

These recommendations are written in JSON format, consistent with the Preservation Action Registries (PAR) data model [8], and published to a Preservica Registry using an API that is consistent with the PAR API definition [8].

B. Executing Processes

Once published, these processes will be automatically executed by Preservica's *Automated Active Digital Preservation* (ADP) feature.

Preservica's architecture allows for individual "mini-services" to be containerised and deployed as consumers of specific messages brokered by a message queue. Specifically, these are implemented as Docker containers, and can be deployed in a scalable manner using a service such as Kubernetes.

As well as allowing for the independent scaling of each mini-service, this deployment model also means that each mini-service can be deployed in an isolated manner, allowing us to avoid resource contention with other parts of the system.

The orchestrator for Automated ADP is one such mini-service, whose function is to watch the Registry for new Recommended Processes, and then query the repository to get a list of Assets that match the criteria in the recommendation. Once this list is generated, it posts a message for each Asset, requesting a re-characterization. These messages are consumed by a separate mini-service, dedicated to performing characterization.

This means that during periods where large numbers of re-characterization processes are requested, we can scale up the number of mini-service instances dedicated to running them. Conversely, once the demand has died down, we can scale back down, meaning that we only use computing and memory resources as we need.

The execution of these processes is explicitly designed to be a background activity that does not necessarily surface to the users of the system. However, it is still useful to be able to track them as they happen, and so each process that is executed is also monitored. This allows us to record general progress updates that detail how far through the list of Assets we are, as well as data and/or process specific error messages (such as forwarding error messages from the characterization tools themselves).

IV. Scale Testing

In order to ensure that this process would be viable, we undertook a program of scalability testing, with a view to replicating the typical scales seen by our cloud systems. This was largely achieved using two distinct testing regimes. The first was "code-

level" integration tests, which gave us tests we could spin up one demand on local development machines, and where we could actually debug into individual processes. The second was to create an actual cloud environment using production level hardware specifications, and populated with large volumes of data.

A. Integration Performance Tests

At the lowest level, we created performance testing at a code level, writing integration tests that configure and deploy the relevant set of mini-services, populate a test database with data, and then trigger background re-characterization processes. For the sake of simplicity, these provide dummy implementations for dependencies like archival storage; only create data that we intend to re-characterize; and use the same input content for each database record.

This means that we are not using them to derive realistic or expected production performance metrics, but they do allow us to quickly run tests with increasing volumes of content to determine where bottlenecks may emerge.

They prove exceptionally useful in replicating issues uncovered in the more realistic test scenarios, allowing us to diagnose those issues, and have some confidence that we have actually resolved them.

B. Production Like Test System

The second and main testing mechanism we used was to create an actual cloud environment using production level hardware specifications, and populated with large volumes of data. From here we could publish realistic recommendations and allow the system to run through re-characterizations in a real world scenario.

This system was loaded with close to 345,000 pieces of content in 763 different file formats (plus around 29,000 "unidentified" formats). As with our production systems, this was heavily weighted to common formats, with over 53,000 JPEG 1.01 files and over 24,000 Word 97-2003 files. The top 10 file formats accounted for over 55% of all content.

By combining formats in our recommendations, we could create processes that would target an arbitrary number of assets to re-characterize. We published a series of recommendations, triggering re-characterization processes on increasing numbers of assets, from tens at a time up to just under 100,000. By querying the monitoring API and underlying database, we could calculate the rate at which these re-characterizations were performed.

For this initial round of testing, we did not perform any scaling of any of the mini-services involved, so at any given time, there was only instance of a mini-service running.

C. Results

The predominant finding from this was that over increasing scales, the rate at which we were able to process re-characterizations did hold relatively constant.

In the majority of test cases run, the rate, as measured by the overall running time of the process divided by the number of assets processed, was less than 1 second per asset. (varying between around 0.1 and 0.7, but averaging around 0.25). In the final iteration of the code, this held true up to the largest dataset we tested, which was in excess of 96,000 assets being re-characterized in a single process.

This is not to say that characterization of any given asset took less than 1 second, since, even though there was only one instance of a mini-service, internally it runs up to 8 threads simultaneously, so 8 assets, each taking 8 seconds to process would still result in a rate of 1 asset per second.

This parallelism benefit could in fact been seen in one of the smallest tests we ran where just 17 assets were being processed. The rate for this test was 5.6s per asset. On closer examination we determined that this was essentially a "small sample effect"; one of the test files was orders of magnitude larger than the others (around 3.5GB), and the overall process time was dominated by the retrieval of this content.

At this rate of less than 1s per asset, processing of up to around 100,000 assets will run for approximately a day, which is well within the comfort zone of being able to generally assume full system uptime.

D. Issues Uncovered

The first issue we encountered was at around 15,000 assets being processed. The rate jumped from less than second per asset to over 3s per asset. The process reported a lot of errors that were ultimately due to calls to Third Party characterization tools being timed out (i.e. cancelled when they took

>30s to return). Although this initially seemed like it might be to do with overwhelming the mini-services, the actual root cause was discovered to be a scalability limit in our "working area" shared storage.

In order to run characterization tools against the content in the repository, we take a copy of the content from its archival storage location (in this case an AWS S3 bucket) and place it in some storage that is accessible to all mini-services (in this case an AWS EFS drive). The throughput on the EFS is throttled by default, giving you an allowance that you use when performing reads or writes to disk, and which replenishes over time when no activity is taking place. At this scale, we were using up all of the allowance without it being able to recover. At that point, all I/O operations became slower than we were able to tolerate. This is relatively trivial to fix, albeit at increased service cost.

The second issue was also due to the same EFS system, or at least, how it was "mounted" in the mini-services, and hit at around 25,000 assets. To reduce network costs, each client connecting to the EFS drive maintains a local cache of what is on the drive. In real time terms, these caches are short-lived and so once a client has written content to the drive, all other clients will "see" the content very shortly thereafter. In our case however, the messaging between mini-services was quick enough that the code that should use the content was trying to read it before its cache updated, then compounding this issue by storing this "not found" result in cache for long enough that eventually the process was timed out. Whilst this was likely happening on smaller scale tests, only at this point did it cause an appreciable impact on our results.

The final major issue that we encountered was to do with the way the processes were being monitored. This presented as an inelastic threshold in our testing. The rate of processing held constant up to around 80,000 assets, at which point, the Automated ADP orchestrator service became very unstable, restarting frequently, causing monitoring to go awry and process requests to be re-sent multiple times.

The limit here was essentially that each time a process completed, we were attempting to update the monitoring information to indicate how far through the process we were. In doing this, we were retrieving a list of the requests, then aggregating them by their process status so that we could update these numbers in the database. There were two issues with this, the first is that at some point, the volume of data contained in the list of requests became large enough that the SQL query to retrieve it would take a long time to complete. The second is that because we were operating 8 processes in parallel, we would often have 8 threads making that call simultaneously. This combination caused contention for database resources, which ultimately cascaded into a series of timeouts and errors.

The issue of counting lots of simultaneous updates in a transactional manner is a common problem in large scale systems, and the general solution is to reduce the number of times you actually update progress, caching all the updates in memory in between. The updates in question here were purely for monitoring, and in large scale processes it is generally acceptable to see updates at longer discrete intervals, so we were able to solve this issue by a combination of performing the status aggregation in the database query (thus reducing the volume of data we needed to transfer), and by only updating periodically (thus reducing the number of database calls we needed to make).

E. Testing Limitations

The system we ran our testing on was configured as a production system would be, with the same hardware specifications, so the direct performance results should be comparable. However, we were limited in how far we could fully replicate a production system in the time available.

At over 345,000 pieces of content, this system was larger than a number of our production systems, but at least an order of magnitude smaller than the largest systems we have. The data also contained many more duplicated items than we would reasonably expect a production system to contain. This introduces some uncertainty into validity of the process. Some data will cause issues with third party tools that other data in the same identified format will not, possibly due to the use of features of that format, or just whether it is valid content. If our dataset contains lots of replicas of problematic data, then this might mean that our measured rate is over-estimating how long a truly heterogeneous data set of the same size would take. Similarly, if it is replicating more "clean" data than would exist in a

truly heterogeneous set, then we might be underestimating how quickly we would process that set.

The final limitation we have identified is that our test system was configured to be single-tenant, whereas many of our biggest systems are multi-tenant. The system is designed to run processes on a tenant by tenant basis, which means that the number of tenants in the same system should be irrelevant, however this set of tests was not designed to explicitly verify this.

V. Into Production

Following on from this successful scale testing, we have started to roll this feature out into live production systems. At the time of writing, this has been limited to around 10 recommendations, across two production systems, reaching scales of up to around 15,000 assets being re-characterized in a single process. Taking the same rate measurements as we did for the testing processes, our performance has been between 0.2 and 0.25 seconds per asset, which is perfectly in line with the results from the test systems.

We will be continuing to enable this feature on more systems, and publish more recommendations over the coming months.

VI. Conclusions

We have reported on a large-scale issue that that affects the users of Preservica's cloud systems, namely that it likely that some proportion of the content they have ingested has outdated characterization information. The result of this is that we are likely to make poorly-informed decisions as to how to treat this content; particularly we may repeatedly attempt to perform processes, such as rendering or migration, that have no prospect of success, which will harm our efforts to preserve information efficiently.

We have discussed the general approach we have taken to implement functionality within Preservica to address this issue. We are allowing Digital Preservation experts to publish machine actionable recommendations for re-characterization processes that should be run, and then automatically executing those within a scalable architecture.

It is noted that for now, assessing updates to the PRONOM dataset as they are formally released is a task that is carried out manually by digital preservation experts using the tools and approaches created in-house for the task.

The additional workload this requires will scale with the number of file formats in the PRONOM database, and the number of types of underlying digital content these represent. This is independent of the volume of content in any given system. For context of the current scalability of this task, PRONOM updates are comparatively infrequent (2 to 3 per year), which limits the frequency at which such analysis has to be performed, and although the sizes of updates vary, they are comparatively small, affecting tens to a few hundred formats each. This makes it possible for a single individual to assume responsibility for this task at each update.

This work is currently being performed by Preservica staff as part of our ongoing digital preservation activities. The output from the analysis is being published for the benefit of the community at

Since the types of data changes that may warrant a re-identification recommendation have so far proven to be relatively systematic, it would likely be possible to augment this process through further automation, perhaps through machine-assisted or machine-learning-based approaches, however exploring these approaches is beyond the scope of this paper.

Over time, it may be necessary to partition this workload so that experts in different types of digital content are responsible for making recommendations related to their expertise (e.g. one expert assessing the impact on images, whilst another assesses the impact on Audio-Video content). Again, this is beyond the scope of this paper.

The mechanism of allowing digital preservation experts to publish recommendations written in a PAR-like data model, and using a PAR-like API means that it should be possible to extend PAR to encompass this in the future. This would enable experts and practitioners from across the digital preservation community to publish their own advice, and access that of others, in a machine actionable way. This would further extend the benefit of this work and enhance knowledge sharing for the entire community and not just Preservica users.

We have presented a description of the testing we have undertaken to validate that this approach will indeed be able to meet the scale of the challenge, summarizing the key results of that testing, and highlighting the key issues uncovered. We have also reported some initial confirmation from production implementation of this feature that our test findings are in line with the performance we are able to achieve on live systems.

We clearly have further work to do in rolling this feature out more generally across our cloud estate, and this work is currently in progress.

The next step in our Automated ADP feature implementation is to enable similar automation of expert derived recommendations around migration functionality. Much of the testing we have already performed will be valid for this also as much of the triggering and monitoring mechanisms are shared. Typically however, migration itself is a more compute and memory intensive process then characterization, so there are still outstanding questions of scaling these processes to answer.

VII. References

[1] M. Hutchins, "Testing Software Tools of Potential Interest for Digital Preservation Activities at the National Library of Australia," National Library of Australia, Canberra, 2012.

[2] T. Gollins, "Parsimonious Preservation: Preventing Pointless Processes!," in *Online Information 2009*, Online, 2009.

[3] The National Archives, "Digital Preservation Workflows > 2. Ingest," The National Archives, [Online]. Available: https://web.archive.org/web/20221202160658/https://www.nationalarchives.gov.uk/archives-sector/projects-and-programmes/plugged-in-powered-up/digital-preservation-workflows/2-ingest/. [Accessed 02 12 2022].

[4] Y. Tunnat, "Sherlock Carriage – PRONOM's blind spot on (some) PDFs from 2010 to 2014," 25 July 2018. [Online]. Available: https://openpreservation.org/blogs/sherlock-carriage-pronoms-blind-spot-on-some-pdfs-from-2010-to-2014/. [Accessed 19 June 2023].

[5] Y. Tunnat and M. Lindlar, "Time-travel with PRONOM: The 4th dimension of DROID," in *iPres 2019 Conference*, Glasgow, 2019.

[6] D. Clipsham, "PRONOM & Preservica's Auto-Preservation functionality," in *iPres 2022 Conference*, Glasgow, 2022.

[7] J. O'Sullivan and J. Tilbury, "Using preservation action registries to automate digital preservation," *Journal of Digital Media Management,* vol. 9, no. 3, pp. 240-252, 2021.

[8] Open Preservation Foundation, "PAR Overview," [Online]. Available: https://parcore.org/. [Accessed 20 06 2023].

BE CAREFUL WHAT YOU CAMPAIGN FOR

How formal organization practice may negatively impact adaptability aspects of preservation

Daniel Steinmeier
KBNL National library of the Netherlands
The Netherlands
Daniel.steinmeier@kb.nl

Abstract – Digital preservationists often struggle using their expert knowledge to create change within their own organization. Because of this, they might need to resort to campaigning for decision-making authority. Why is this? Memory institutions are used to adhering to standards and rules. Rules and regulations are beneficial for stability and trustworthiness. But too much focus on rules may create organizational rigidity which negatively impacts adaptability. Adaptability is a major goal for preservation so how could we create more room for this? An important part of adaptability is organizational learning. In order to facilitate learning we must understand which aspects of organizational practice negatively affect it. For example, avoiding discussion of mistakes is an important barrier to learning. If an organization prioritizes learning this can have a positive impact on the motivation of employees. Practitioners may feel more in control when they understand how to use theories of organizational learning to further implementation of preservation principles. More room for learning within the organization might also benefit the field of preservation itself through enhanced knowledge of what works and what doesn't.

Keywords – Organizational theory, learning organization, adaptability, stability
Conference Topics – From theory to practice.

I. INTRODUCTION

In my ten years of experience with preservation, one of the most striking and enduring aspects of this line of work is how much preservation practitioners know about their area of expertise and at the same time how difficult it is to use this knowledge to get preservation requirements implemented within the own organization. When I first read the article 'What's wrong with Digital Stewardship?' three years ago I was amazed by how similar the findings were to this first impression of mine. I always encourage people to read this article because I cannot do justice to it with a short summary. But what I learned from this article is that it is hard to implement a holistic model with a long-term focus on adaptability in a hierarchical organization that values short term measurable results and separate roles and responsibilities [1].

The article neatly describes some significant organizational factors that negatively affect implementation of preservation principles according to practitioners. For instance, hierarchical structures that disempower experts from taking part in decision making, leaving them no other option than campaigning for authority. In this article I will describe where we might look for improvement. An important part will be analysis of the organizational level and the implicit rules and restrictions that come with a certain organizational practice. Arguably the focus within organizations on rules and policies is what makes the stability goals of preservation easy to relate to. But how does this affect the goal of adaptability that is necessary as well for the model to actually work? And is the field of preservation itself keen enough to adapt their principles to new insights and changing circumstances? As I will argue in this article, paying attention to lessons from the field of organizational learning might help create a more complete implementation of preservation functions within the organization. The process might in its turn also benefit the field of preservation itself through enhanced knowledge of what works and what doesn't.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

II. The conflicting values within digital preservation: exploration vs. exploitation

Memory institutions usually do not seem to be daunted by rules and regulations. Coming from a library background I would argue the huge amount of rules related to title descriptions for cataloguing publications is a case in point. The requirements for Trustworthy Digital Repositories (TDR, also known as ISO-16363) about creating policies and fixed procedures for handling objects during the digital lifecycle should not feel like too much of a stretch. A preference for rules and policies is one of the traits that characterize people drawn towards the public sector (among other characteristics, the most surprising one is probably self-sacrifice!) according to the concept of Public Service Motivation [2]. Rules, policies and documented procedures are supposed to prevent ad-hoc actions that could lead to unpredictable decisions that might endanger digital objects or the trustworthiness of the repository. The goal is to create more stability through bureaucratic methods. However, stability is also the opposite of flexibility. Organizations in modern times need to be flexible to be able to keep up with technological change. This is especially relevant within the field of digital preservation that was conceived to a large extent with the goal of countering obsolescence by staying up to date.

Exploration of new avenues and exploitation of existing knowledge are contradictory processes within organizations that need to be balanced out if the aim is to profit from both. Exploration should not be constrained by existing rules while exploitation should benefit from new knowledge that is generated through exploration [3]. The same dual focus can be seen within the field of preservation since the OAIS-model is describing functions for exploration of new technological developments and changed user demands, as well as functions for creating stability and use of existing knowledge through documented procedures. If for no other reason than sheer familiarity, we would expect memory institutions to feel more at home with the stability goals than with the flexibility goals of digital preservation. As a tentative suggestion that this is indeed the case we may look for example at the NDSA survey of 2021 where among all the functional areas listed as relevant to preservation no explorative (informal learning-oriented) functions such as Preservation Watch are present [4]. The areas that are listed involve technical implementation, planning and policy writing which implies exploitation, consolidation and streamlining of existing knowledge and practice.

In the requirements for certification, we also see a heavy focus on stability. This can be seen in requirements that describe the need for fixing organizational procedures by way of integration and documentation. Both ISO-16363 and CoreTrustSeal [5] refer to documented processes as proof that ad-hoc decisions are minimized. CoreTrustSeal specifically has a whole requirement (R11) dedicated to this. As can be seen however from literature on the effectiveness of process management, implementing fixed processes can have negative consequences for innovation and flexibility. Within a stable and predictable environment process management can increase efficiency and therefore benefit the organization. However in an innovative and changing environment it can negatively impact results because organizational learning and creativity is hampered by processes that are based on exploitation of existing knowledge. This focus can even lead to resistance to change [6]. The reliance on documentation as evidence of trustworthiness is also motivated by the fact that this makes the auditing process more objectively verifiable. However, risk is socially constructed. The creators of the standards, the auditors and practitioners can have different opinions on what the most relevant risks are and what the best way of mitigating these risks is. These differences are rooted in the different stakes persons have through their various roles in the process of certification. For example, from a standards perspective it is useful to have written proof of continuity in the form of a succession plan but practitioners may have doubts that this method is effective in countering risks to continuity [7].

Too much focus on rules and regulations not only makes us less flexible but the ideal of finding universal solutions also ignores the fact that knowing the context and the specific cultures of organizations is important if we want to implement solutions that fit the environment. Organizational culture should not be treated as something that hinders ideal implementations but rather as something that needs to be understood in order to create room for diversity. By understanding the complexity of human behavior, for instance in decision making, solutions

can be found that better match real-world situations [8].

If we only focus on rules based on existing knowledge this will not impede aspects of preservation that are supposed to provide stability. But it may have negative effects on those aspects of preservation that imply the existence of processes of cyclical, informal learning and improvement. An example of this is Designated Communitiy monitoring. The goal of monitoring the Designated Community is to signal when changes are needed to information or services through regular gathering of non-expert, informal information. As can be seen from the literature, there is still a dearth of information on how to implement this concept, while it is foundational within the model. The concept itself also reflects conflicting values in the sense that public institutes aim to serve a broad community while the model requires being specific about what is done for whom, implying a more exclusionary definition. Furthermore, this function is aimed at adapting to the needs of future users as well, not only to the requirements of current users [9]. It seems we need to learn more about the concept as well as about how to apply it. Using only existing knowledge to justify our policies and practice, it is not likely that we will achieve the goal of catering for future requirements. We cannot just follow a training or ask a colleague. And even the familiar method of doing a survey would not suffice since this only targets our existing user base.

If we want to adapt to new developments we need to learn how to innovate and learn, not just on a personal level but also at the level of the organization. We got the stability aspects of digital preservation in clear focus through policies and standards, but what about the adaptability aspects? The fact that organizations struggle with more explorative concepts of the model, like the idea of the designated community, may be a hint that these aspects do not work smoothly with formal organization practices still current within memory institutions.

III. Implementing a holistic model in a formal organization

It is not only explorative concepts that might be relatively unfamiliar to memory institutions but also another important aspect of the OAIS-model: the fact that it is a holistic model that affects the inner workings of the whole organization. In 'What's wrong with digital stewardship' this is described as a major stumbling block [10]. When we understand what characterizes the formal organization we can better understand why this is so. The idea of the formal organization was conceived of from the perspective of efficiency within a factory work line so as a classic example, think of a factory a 100 years ago. It is clear what the product should be, who works on which parts, how the parts should be assembled and how many products can be manufactured per day. The production process as well as the distribution of associated tasks are determined by top management because all necessary knowledge about the product and about efficient processes is concentrated at the top. The employees who work on different parts do not have to understand how the whole production line works. They just need to ensure that they can perform their own limited task within the bounds of the production standards set by the top management. Of course, working in a factory can be very different in practice, but this is to outline the extreme end of the spectrum by describing the workings of a very formal organization. In this type of organization, employees who do not belong to the top are expected to be performing simple, predetermined tasks that form just a small part of the whole. Employees are rewarded on the basis of achieved results [11]. In lots of ways aspects of the formal organization - such as decision making at the top – are still prevalent in modern organizations, and indeed in memory institutions. It should not be a surprise that the holistic view of preservation activities across the organization and the shared responsibility implied by this view should not sit well with people who are used to having clearly separated roles and responsibilities within a chain of command. This is borne out by the conclusions in 'What's wrong with Digital Stewardship' and most aptly formulated in the chapter title “Hierarchical organizations exacerbate stewards’ lack of authority”. Also other stumbling blocks mentioned in this report can be linked with the workings of the formal organization. For instance lack of long-term commitment and structural funding is linked to the fact that organizations make decisions based on financial benefits in the short term, such as project funding.[12] A focus on short term, measurable results is also a clear feature of the formal organization. As mentioned earlier, having separated

and clearly defined roles is a feature of the formal organization and a means for achieving efficiency but it hinders coordination of preservation activities across the organization.

IV. Laying the groundwork for adaptability: organizational learning

What should be clear from the above is that aspects of the formal organization seem to agree with the formal aspects of the OAIS model but not so much with preservation principles aimed at adaptability. One of the aspects that suffers most from the rigidity of the formal organization is learning because workers are trained not to ask questions or question authority [13]. If we only consolidate existing practice we are not discovering new ways of doing things. Theories of organizational learning provide solutions on how to create better conditions for learning to take place. Learning itself is an important part of preservation but creating the conditions necessary for learning should also improve other aspects that hinder implementation such as hierarchical decision making.

So what does the ultimate learning organization look like? The idea of the learning organization in its most extreme form (think of a small startup) is in many ways the complete opposite of the formal organization. In order to be more open to learning, it is necessary not to determine everything in advance, to give employees autonomy, to show initiative, to accept mistakes, to flatten the hierarchy and to gradually learn what works by trying new things. Transparency is paramount and critical thinking is seen as a crucial skill to improve things. Employees are rewarded for having the right attitude, not for successfully performing planned actions [14].

The differences between these two organizational views can best be illustrated with the concept of collaboration. From the idea of the formal organization, being a good employee means that you stay within the boundaries of your role as much as possible and not try to do things that fall outside your jurisdiction because this hinders efficiency. Failing to do this can be perceived as meddling and lack of trust (even if people might be too diplomatic to say this out loud).

From the perspective of the learning organization, however, it is important that there is overlap (at the cost of efficiency) because in this way new ideas may come up. This process is called creative interference in the literature. Within a learning organization, it is considered beneficial when people are working on the same thing from different perspectives. The contribution of group members might be based not only on their professional knowledge but also on their personal knowledge. People can have useful information that is not part of their job description but contributes to solving a problem in a way that has not been tried before [15]. It is important to understand that desirable outcomes that are typical for a learning organization, such as creativity, autonomy and innovation, cannot be achieved by using the methods and goals of the formal organization such as efficiency, separation of tasks and planning. If we want both, we need the right balance of formality and learning, but without the two sets of methods and expected outcomes getting mixed up. Knowledge and awareness of the differences between these opposite orientations is the first step. Organizational learning can contribute to achieving preservation goals because it will enhance creativity and innovation. If we only focus on rules and regulations we miss the learning-focused aspects of preservation that will help us adapt to new developments in the long term.

V. What we should unlearn

These insights on what stimulates innovation have been around for some time and might sound vaguely familiar to people acquainted with agile software development principles. So why is it so difficult to put this knowledge into practice? Though there might not be one simple answer to this question, I think this can partly be explained by realizing the formal way of doing things is so enmeshed with things we value and things we are used to, like our expertise and our way of communicating.

The latter point will become clearer by zooming in on the work of Chris Argyris whose professional output has been significant in the field of organizational learning. In his work, he stressed the importance of looking at the underlying values within a process of problem solving that effectively prevent change from happening. The organizational process where certain types of solutions are automatically selected creates the effect that new solutions, after a while, will start looking very much like the old

problems they were meant to solve. Underlying values steer behavior within a problem-solving context. When people realize that their values in dealing with problems is what is creating failure this is called 'double-loop learning' and is important for organizational learning to happen [16]. If we want to translate this idea into something recognizable within the field of preservation, we could think of the ingest process. According to the OAIS model, quality analysis should be part of the ingest process. The goal is to safeguard quality. However, this step might take time during which the content isn't being preserved. A single-loop solution would be to try and speed up the process. Double-loop learning would be to question the underlying values of the solution and try to provide alternatives, such as ingesting first and then doing quality analysis as was proposed by the authors of the "Minimal Effort Ingest"-Ipres-paper [17].

According to Argyris, the way we communicate can prevent double-loop learning from happening. A diplomatic way of dealing with mistakes and criticism fits the formal organization. One can think of face-saving actions after dubitable decisions and giving reassurance to people to protect the trust they have in the chain of command. This type of diplomacy is protective behavior that leads to a reduction in transparency and is therefore a barrier to learning. From the perspective of a learning organization, conflict avoidant behavior gets in the way of detecting mistakes and learning from them. It is also something that becomes automatic behavior which means people aren't even aware they are doing this [18]. In preservation this could happen if we do things because the guidelines state this as a necessity or because it is considered a best practice. This means the guidelines or best practices themselves aren't open to questioning anymore, only the solutions based on them are.

This problem can be solved by actively facilitating critical thinking within the organization. For example, by making statements based on facts that can be verified or tested by others because the same information is made available to everyone. This also requires a certain attitude towards questions in the sense that asking for verifiable facts is not perceived as distrust but is rewarded as an attempt to stimulate open communication. Reciprocity and flexibility are important here: we must be open to adapting our ideas on the basis of verifiable, factual arguments. Learning from mistakes is an important part of this. In practice it might be hard for preservationists to openly discuss mistakes because trustworthiness is one of the core concepts in the field of preservation. Making mistakes might mean data loss which shouldn't happen in trustworthy digital repositories, right? So how should we improve our attitude towards mistakes? One of the first steps to change is making undiscussable things discussible. The second step is to confront any threat or embarrassment that might result from this, instead of avoiding it [19]. In the case of preservation this might mean talking about which decisions actually resulted in data loss. In this way the organization can use this information to come up with alternative solutions. If we do the opposite and automatically avoid embarrassment we might stick to policies and solutions even though they are not effective and maybe never even realize that this is the case. If we want to optimize processes of learning we also need to take the human factor into account. Motivation is an important part of learning so this is what we turn to next.

VI. The human factor

As is very clearly stated in 'What's wrong with digital stewardship', constant campaigning for decision making authority is said to lead to burn-out and frustration. The situation of the digital steward, according to this report, is often one of autonomy without authority [20]. That is to say, people are free to explore options but they do not have the authority to implement solutions. There are indications that preservationists are not alone in this. Research on academic libraries for example, points out that library staff in general may experience lower morale due to status differences, lack of participation in decision-making and silo-ed communication. These are all features of the formal organization and point to organizational barriers leading to personal difficulties [21]. Having experts outside the management team but not involving them in decision making seems like a clear case of mixing up elements of the formal and the learning organization in a way that is counterproductive. As stated before, in a very formal organization both expertise and authority are vested in the top management levels of the organization. The situation as described in the report is one where the expertise has trickled down to the lower levels of the hierarchy while the decision

making authority didn't follow suit. This leaves the preservationist no other option than communicating and campaigning to the point of exhaustion. This takes up time that isn't spent on implementation, testing what works and learning new things about preservation. Bureaucracy has an impact on creativity both by centralizing decision-making and by providing rules (formalization). There are indications that centralization, more than formalization, is an environmental factor that deactivates creative behavior in learning-oriented people [22]. If this is correct, then the lack of authority should be a priority to solve. The stability aspect of preservation needs rules and procedures and therefore it is good to know that formalization doesn't necessarily hinder creativity. But both stability and adaptability are negatively impacted by centralized decision making, respectively through siloed communication and undermined creativity. Therefore focusing on decentralization and empowerment should benefit both exploration and exploitation goals of preservation.

If we want to understand what organizations can do to empower employees, we can turn to self-determination theory. This is a broad framework for the study of motivation and personality [23]. This overarching concept consists of a number of mini-theories, one of which is Basic Needs Theory. The central tenet of this theory is that people have a basic need for autonomy, competence and relatedness. This means people need to feel in control over their actions and behaviors, people need to feel a sense of mastery over their environment and people need to feel a sense of bonding with other people. As stated before, ideas about the formal organisation started out from an industrial perspective which led to fragmentation and simplification of tasks and external control over these tasks. It should not be difficult to see how external control is thwarting the need for autonomy and how simplification and fragmentation might have a negative impact on the need for competence. Formalization as a way to enhance organizational control has been said to lead to alienation within the public sector. This is why learning organizations focus on improving job satisfaction by empowering workers, making them part of the decision-making process and prioritizing learning [24]. The organization benefits from individual learning while the need for self-development that employees might have will also result in commitment to the organization through shared goals of learning [25]. Especially people within the organization who are high in Need for Cognition, that is to say, people who enjoy effortful cognitive activities, will benefit from an organizational culture that rewards innovation and creativity. This will lead to greater activation of creativity which is important for organizations to adapt [26].

An important part of improving the motivation process within the organization is creating awareness about motivating styles. This concept centers on how employers motivate their employees. Motivating style can range from controlling to autonomy support. A controlling style means being prescriptive and being insistent on what employees should think and do. The autonomy supportive style, on the other hand, is one where respect for the perspective, input and initiatives of employees is salient. For instance, by explicitly asking the perspective of employees, providing rationales for decisions and using a non-pressuring, informational communication style, among other things. The management of an organization needs to understand what their motivating style is and how to change this style if necessary. A controlling motivating style can be improved by teaching supervisors the principles of autonomy support. The autonomy supportive style leads to conditions that support and satisfy the basic needs of autonomy, competence and relatedness [27]. Basic need satisfaction leads to autonomous motivation which has been linked to aspects of well-being, including commitment and work performance [28].

Through the above suggestions of how learning organizations improve motivation, we get a glimpse of what could help remedy the situation of the authority-deprived digital steward. The downside is that we will probably need to campaign for this as well…

VII. What we can control: choosing to react differently

Instead of campaigning for implementation of preservation solutions, we could therefore be strategic and campaign for implementation of principles inspired by learning organizations in the hope of being granted more discretion in handling preservation issues. But this still implies external

control of our goals. What can we do ourselves to feel we are making progress in our area of expertise?

If indeed it is the case that digital preservation practitioners have autonomy without authority, then at least we can put the autonomy to good use. As mentioned above autonomy is one of the basic needs according to Basic Needs Theory so to be able to explore new solutions according to our interests is in and of itself a valuable asset. Setting up small experiments, even thought experiments, could help maintain a better balance between stability and adaptability. Previously mentioned pitfalls of mixing up formal and learning methods may also happen on the individual level so it is important to be aware of our own reactions and problem-solving techniques if we want change. Say for instance we want to problem-solve the previously mentioned lack of decision making available to the digital steward by petitioning management to enforce clearer roles and responsibilities. Given the above we should realize that this is a solution from the formal organization which in the long run will not enhance learning, intrinsic motivation and flexibility. After all, more authority for the preservation practitioner does not mean more egalitarian decision-making processes for everyone. Instead, we could consciously choose to adopt an alternative solution taken from theories about organizational learning. For instance, raising awareness about the benefits of autonomy support. In an indirect way this could solve the problem as well, but without negatively impacting flexibility and motivation. It is important to realize that the things we campaign for can contribute to stability goals or adaptability goals but both types of goals need different things and are opposing values that need to be balanced carefully. When we decide to take action, we can consciously choose to use a method that contributes to a better balance between these two types of goals.

If we want to help create a learning climate within the organization (and also in the broader network of memory institutions), it is important to be aware of our own communication style. Instead of presenting preservation requirements as strict rules we could open them up for questioning and communicate improvements in a non-pressuring way by providing rationales and options where possible. By providing autonomy support we can appeal to intrinsic motivation. If we are aware of our own communication and motivation style during campaigning and make sure that this provides room for other perspectives, then we can embody learning-oriented values as an example for others.

VIII. Conclusion

Despite all the knowledge we might have on how to preserve digital objects in the long term, failing to get the message across within our own organization has been the metaphorical elephant in the room. Hierarchical organizations have a focus on rules and procedures that is partially matched by preservation requirements aimed at providing stability and trustworthiness. By focusing too much on stability, however, explorative aspects of preservation might suffer, endangering the other important preservation goal of adapting to new developments. Part of the reason for the stability focus is the fact that formal organizations are not tailored to holistic, bottom-up, informal learning processes. Centralized decision making is an important barrier in this respect. Lessons from theories of organizational learning can help us understand what can be done to stimulate innovation, creativity and learning within our own organization. This can be difficult as it means also changing the ingrained ways of communicating and problem solving that we have come to associate with being an expert. Focusing on methods for organizational learning might benefit those aspects of preservation that are focused on learning and improvement, like the concept of the Designated Community. This might also help the field of preservation itself to adapt and improve its methods through the infusion of new ideas.

1. REFERENCES

[1] K. Blumenthal, et al, "What's wrong with digital stewardship: evaluating the organization of digital preservation programs from practitioners' perspectives", in *Journal of Contemporary Archival Studies* vol. 7, no.1, pp. 1-22, 2020.

[2] D. de Gennaro, "Transformational leadership for public service motivation", in *Journal of Economic and Administrative Sciences*, vol. 35, no. 1, pp. 5-15, 2018.

[3] M. Benner and M. Tushman, "Exploitation, exploration, and process management: The productivity dilemma revisited." *Academy of management review* vol. 28, no. 2, pp. 238-256, 2003.

[4] 2021 Staffing Survey Report - an NDSA report, https://osf.io/2rb7k

[5] CoreTrustSeal requirements 2023-2025, https://zenodo.org/record/7051012

[6] M. Benner and M. Tushman, "Exploitation, exploration, and process management: The productivity dilemma revisited."

Academy of management review vol. 28, no. 2, pp. 238-256, 2003.

[7] R. Frank, "The social construction of risk in digital preservation", Journal of the Association for Information Science and Technology vol. 71, no. 4, pp. 474-484, 2020.

[8] F. Foscarini and G. Oliver, "The information culture challenge: moving beyond OAIS", online available at: https://digitalpreservationchallenges.files.wordpress.com/2012/09/foscarini.pdf

[9] R. Frank and L. Rothfritz, "Designated Community: uncertainty and risk", *Journal of Documentation,* ahead-of-print, 2022.

[10] K. Blumenthal, et al, "What's wrong with digital stewardship: evaluating the organization of digital preservation programs from practitioners' perspectives", in *Journal of Contemporary Archival Studies* vol. 7, no.1, pp. 1-22, 2020.

[11] G. Dosi, L. Marengo, and M. Virgillito. "Hierarchies, knowledge, and power inside organizations." *Strategy Science, vol.* 6, no. 4, pp. 371-384, 202.

[12] K. Blumenthal, et al, "What's wrong with digital stewardship: evaluating the organization of digital preservation programs from practitioners' perspectives", in *Journal of Contemporary Archival Studies* vol. 7, no.1, pp. 1-22, 2020.

[13] J. Sarros, et al., "Work alienation and organizational leadership", *British Journal of Management,* vol. 13, no. 4, pp. 285-304, 2002.

[14] F. Walumbwa, A. Christensen, and F. Hailey. "Authentic leadership and the knowledge economy: Sustaining motivation and trust among knowledge workers." *Organizational dynamics*, vol. 40, no. 2, pp-110-118, 2011.

[15] T. Lumley, "Complexity and the "Learning organization"", *Complexity*, vol. 2, no. 5, pp-14-22, 1997.

[16] C. Argyris, "Initiating change that perseveres", *American Behavioral Scientist*, vol. 40, no. 3, pp. 299-309, 1997.

[17] B. Jurik, A. Blekinge, and K. Christiansen. "Minimal Effort Ingest", iPRES 2015.

[18] C. Argyris, "Initiating change that perseveres", *American Behavioral Scientist*, vol. 40, no. 3, pp. 299-309, 1997.

[19] Ibidem.

[20] K. Blumenthal, et al, "What's wrong with digital stewardship: evaluating the organization of digital preservation programs from practitioners' perspectives", in *Journal of Contemporary Archival Studies* vol. 7, no.1, pp. 1-22, 2020.

[21] Glusker, Ann, et al. ""Viewed as equals": The impacts of library organizational cultures and management on library staff morale." in *Journal of Library Administration* vol. 62, no.2, pp. 153-189, 2022.

[22] G. Hirst, et al., "How does bureaucracy impact individual creativity? A cross-level investigation of team contextual influences on goal orientation–creativity relationships", *Academy of management journal*, vol. 54, no. 3, pp. 624-641, 2011.

[23] Self-determination theory, https://selfdeterminationtheory.org/theory

[24] K. Karyotakis and V. Moustakis. "Organizational factors, organizational culture, job satisfaction and entrepreneurial orientation in public administration", *The European Journal of Applied Economics*, vol. 13, no. 1, pp. 47-59, 2016

[25] D. Jamali, G. Khoury, and H. Sahyoun. "From bureaucratic organizations to learning organizations: An evolutionary roadmap", *The Learning Organization*, vol. 13, no. 4, pp. 337-352, 2006.

[26] H. Madrid and M. Patterson. "Creativity at work as a joint function between openness to experience, need for cognition and organizational fairness", *Learning and Individual Differences*, vol. 51, pp. 409-416, 2016.

[27] J. Reeve, "Giving and summoning autonomy support in hierarchical relationships", *Social and Personality Psychology Compass*, vol. 9, no. 8, pp. 406-418, 2015.

[28] S. Trépanier, et al., "On the psychological and motivational processes linking job characteristics to employee functioning: Insights from self-determination theory." *Work & Stress*, vol. 29, no. 3, pp. 286-305, 2015.

IPARO: InterPlanetary Archival Record Object for Decentralized Web Archiving and Replay

Sawood Alam
Internet Archive
USA
sawood@archive.org
0000-0002-8267-3326

Abstract – We proposed a decentralized version tracking system using the existing primitives of IPFS and IPNS. While our description talks primarily about archived web pages, we proposed the concept of IPMT and namespacing so that it can be used in other applications that require versioning, such as a wiki or a collaborative code tracking system. Our proposed system does not rely on any centralized server for archiving or replay of the content. The system continues to allow aggregators to play their role from which both large and small archives can benefit and flourish.

Keywords – IPARO, IPFS, Decentralized Web, DWeb, Web Archiving

Conference Topics – WE'RE ALL IN THIS TOGETHER; DIGITAL ACCESSIBILITY, INCLUSION, AND DIVERSITY

I. Introduction and Background

Web archiving is the practice of temporal versioning and preservation of representations of resources on the web. An archival replay is the practice of the playback of the archived historical representation of web resources while maintaining the essence and fidelity of the web resource. Decentralized content-addressable file systems, such as InterPlanetary File System (IPFS) [1], offer the opportunity to preserve all the historical versions of files stored in them. However, their current implementations lack native support for versioning.

In 2016, our initial work on InterPlanetary Wayback (IPWB) was a successful exploration of the possibilities of web archiving in the early days of IPFS [2-4]. It gained a fair share of visibility in the relevant communities. However, it relied on a local index for the system to operate, which made its operations centralized, despite the archival data being decentralized. We address this shortcoming in this work to make web archiving truly decentralized.

During the IPFS Lab Day event in 2018, we discussed the limitations posed by the centralized index for decentralized web archives and laid out what was needed to address them [5]. We proposed that the InterPlanetary Name System (IPNS) be history-aware, but current implementations only keep the most recent mapping of a URI to its corresponding content hash using a Distributed Hash Table (DHT), leaving prior versions of a resource untracked when the IPNS mapping is updated. We later proposed IPNS-Blockchain (a decentralized blockchain-based approach) [6] and Trillian data store (an append-only federated solution) [7] to eliminate the need of local index. Since neither of these approaches attracted enough momentum to get implemented, we came up with another solution that operates within the existing IPFS primitives.

In this approach we embed references to prior versions of a resource in the new versions, forming a singly linked list for backward traversal. We then use IPNS to enable direct access to the most recent version and traverse the chain from there. Moreover, we propose media types and namespaces in IPFS to make local indexing of archival resources more efficient, a concept that is orthogonal to IPNS-based access. Finally, we describe ways to achieve deduplication by storing slices of data separately that are likely to be repeated many times. Our proposed system, InterPlanetary Archival Record

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

Object (IPARO), can be implemented without the need of a centralized server and archival records can be played back from a client-side system like a web browser using Service Workers [8,32].

Recent developments of web archiving tools [9-11] have made the practice more accessible to individuals pursuing personal web archiving. However, discovery of and access to the archived web content is still limited to large institutions and organizations, while small efforts suffer from disuse. Making large-scale web archival data available to researchers is still a challenge [12-15]. Our proposed system, IPARO, makes the ecosystem more inclusive and accessible to everyone. Archived content preserved in the IPFS network would be discoverable and accessible to everyone, irrespective of the creators of IPAROs.

II. Related Work

Content-addressing is well-established practice, especially, in the context of data shared over a network. Unlike location-addressing, in which Uniform Resource Locators (URLs) are used to find a resource, content-addressing uses technique like hashing to identify corresponding resources. Content on a URL may change over time, but a content-address is derived from the content itself, so it is guaranteed to maintain the integrity and fixity of the content for a given content-address. Content-addressing has traditionally been used in peer-to-peer and decentralized systems like Torrent and Git. More recently, the introduction of InterPlanetary File System (IPFS) [1] has also appreciated content-addressing in its protocol to leverage advantages like peer-to-peer discovery, replication, deduplication, and integrity. In an IPFS network any data can be added, which results in a content identifier (CID), using one of the many supported hashing algorithms. The CID can then be used to retrieve the data by performing a lookup in the peer-to-peer network. Alternatively, an InterPlanetary Name System (IPNS) entry can be added that maps a URL to its corresponding current CID, which would allow performing lookups using the URL, instead of the content digest. IPNS entries are broken into pieces and stored in a distributed hash table, so that there is no single point of failure, and the table does not grow too large on a single node.

While distributed computing, decentralized web, peer-to-peer networks, content-addressing, and web archiving are well explored fields individually, decentralized web archiving is still a fairly new and niche discipline. After the emergence of IPFS, we explored it for decentralized web archiving the first time in 2016 by introducing InterPlanetary Wayback (IPWB) [2-5]. We processed Web ARChive (WARC) files (an ISO standard file format, used to preserve archival data) [16] to go through HTTP response records, split them in headers and payload, store the two pieces, get the corresponding pair of content digests, and create a local index that maps that archived URL and archiving date-time to the corresponding IPFS hashes of the header and payload. This local index is queried during the replay time to retrieve corresponding data from the IPFS network. The header and payload are split (as opposed to storing them as a single record) before storage to leverage deduplication of payload as the headers can be unique even when the payload is the same in consecutive observations of the same or different URLs (primarily due to the presence of the HTTP "Date" header). We also implemented a Service Worker module, Reconstructive, to allow client-side archival replay with minimum rewriting [8]. The biggest limitation of our IPWB system is the need for a local index to keep track of archived URLs and their corresponding IPFS hashes (i.e., CIDs), which makes it partially decentralized.

Webrecorder has embraced IPFS and other decentralized file systems by adding support for them as storage engines. For example, using their ArchiveWeb.page tool, one can share archived data via IPFS. Similarly, their ReplayWeb.page tool allows playback of archived data directly from an IPFS address. However, these tools use IPFS as a file storage system and store entire WARC files (or the emerging WACZ files), as opposed to individual WARC records (as used by our IPWB system). References to such archived bundles are shared using out-of-band channels as there does not provide any built-in mechanisms for the discovery of the archived content in the IPFS network.

Bamboo is a cryptographically secure, distributed, single-writer append-only log that supports transitive partial replication and local deletion of data [17]. This log format creates a tree-

like linked list that has a logarithmic path length between any two entries (from newer to older). Each entry in this format has a link to the previous entry and another link to an older entry to allow long jumps in the chain. Our resilient linked list concept has similar inspiration, but we allow an arbitrary number of links from any node to its prior nodes. Our system also caters to the possibility of incorporating nodes that were left from being part of the chain earlier due to any race conditions or transient discovery issues.

III. Methodology

We create an InterPlanetary Archival Record Object (IPARO) for every archival observation that is intended to be looked up and replayed independently. These objects contain an extensible set of headers followed by the data in any one of the many supported archival formats and optional trailers, stored as IPFS media types (a concept we introduce in section III-A), as shown in Fig. 1. The purpose of headers is to identify the media type of the data, to establish relationships with other objects, to describe interpretation of the data and any trailers, and to hold other associated metadata.

An IPARO is the basic building block of web archiving in IPFS. It is enough to have a working decentralized web archiving system just by storing IPAROs. Anyone can create and store these, while anyone can scan the entire IPFS network data in the future, identify IPAROs from all the stored objects, and interpret them for replay based on the self-contained information. However, this approach is impractical and inefficient. In the coming sections we address these limitations.

A. InterPlanetary Media Types (IPMT)

Media types is not a new concept, but we propose a way to bring it to the IPFS ecosystem as InterPlanetary Media Types (IPMT), which is otherwise an opaque data storage system. In traditional file systems, applications often rely on file name extensions to identify the format or media types of files. In cases where those are absent or are not reliable, applications attempt content sniffing (such as, reading the first few bytes to find signatures of various media types) to interpret the file appropriately. It is worth noting that a media type is not only about the syntactic interpretation of bits, but also the semantics. For example, a JSON file can

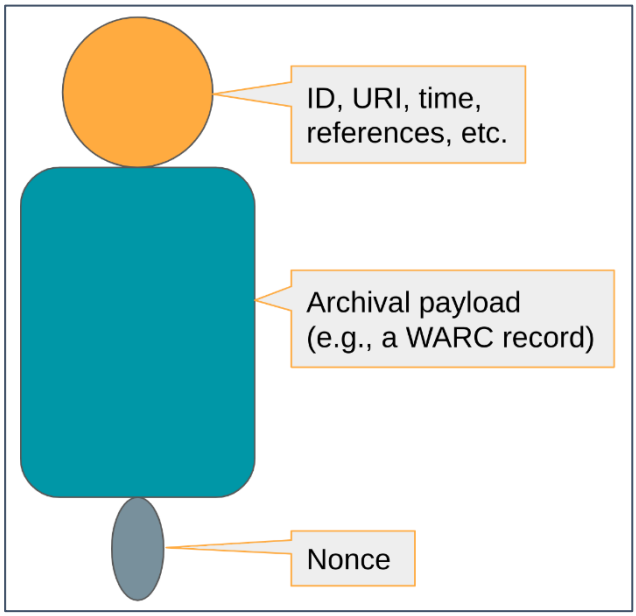


Fig. 1: InterPlanetary Archival Record Object (IPARO) Outline

be interpreted by a JSON parser, but the contents of the file can have configuration information of a system, state of a system, data records of an observation, etc. In Hypertext Transfer Protocol (HTTP) a server can convey the media type (and other clues for the correct interpretation) of a response with the help of response headers on-the-fly. We propose an HTTP-like mechanism for this in IPFS, but in this case the headers should be stored along with data (similar to how WARC records store metadata of each record). However, this means every record will have additional headers that many clients might not understand. One way to solve it is to use a proxy that converts IPMT headers to HTTP headers. Another approach would be to extend the IPFS API to optionally add/remove those headers depending on the need and capabilities of the client. These approaches require changes that might not be desired or feasible. In section III-D we discuss unobtrusive ways to keep the headers and data separate and combine them when needed. The IPFS ecosystem has the concept of InterPlanetary Linked Data (IPLD) [18] which facilitates interoperability among many hash-based systems. It allows external schema definitions to describe the outline of data types. Hence, IPLD can be used to describe IPMTs.

The concept of IPMT is generic so that various use cases and applications can leverage different IPMTs, but in this work we focus on IPMTs that can be interpreted by archival replay systems. Together the set of these IPMTs falls under IPAROs as described

earlier. In sections III-A1 through III-A6 we describe primary IPARO candidate IPMTs, but this list can be extended as more supported archival formats emerge.

1) Memento - We introduce a media type called memento, which in addition to IPMT headers, contains the original HTTP response headers and payload, corresponding HTTP request message (if available), and any additional metadata created by the crawler. This IPMT may also contain references to the corresponding archived versions of all the page requisites as a dependency graph. This media type holds an archival record of only one representation of a resource. Ideally, this can be part of WARC as an "exchange" record, but such a WARC type is not defined in the specifications yet, because the use-case was not realized before.

2) WARC - A Web ARChive (WARC) [16] file is an arbitrary set of records of HTTP request, response, metadata, and various other types of resources (often compressed individually) concatenated together. There are no explicit order, grouping, or limitations defined for the format. Direct access to specific records in a WARC file usually requires an external index (a CDX or CDXJ file [19]). It is possible to store a whole WARC file and optionally its index with many observations of many web resources (such as one or more web pages and all their page requisites) as an IPARO in IPFS, which can then be retrieved and replayed by a client that supports it.

3) WACZ - Web Archive Collection Zipped (WACZ) [20] is a Zip container for one or more WARC files, their CDXJ index, and various other metadata files as a single bundle. It is possible to store WACZ files as IPAROs to be retrieved and replayed by supporting clients.

4) HAR - HTTP Archive (HAR) [21] file format is commonly used in web browsers' developer console for network and performance analysis and debugging of web pages. It is a JSON-based file format that can be used for archival replay directly or after transforming it to something like a WARC file. This is yet another candidate for being an IPARO.

5) Web Bundles - Web Bundles or Web Packaging [31] is an evolving format to deliver multiple HTTP exchanges as a single resource by the primary origin or third-party aggregators/CDNs when signed. This too can be an IPARO.

6) Annotations - IPAROs are supposed to be immutable, so any annotations, contexts, additional metadata, or linking that are realized after the storage of an IPARO must be attached as a separate object (without altering the existing IPARO). An annotation is not necessarily one of the archival formats, but it is of interest in the context of archival replay, so it would be useful to include it in the set of IPMTs that fall under IPARO. The exact format details are yet to be determined, but semantics can be borrowed from existing specifications and practices related to annotations [22].

B. Immutable Linked List

Web archives usually archive each web resource in their collection numerous times over time. As a result, they can report a list of all the observations (also known as a TimeMap [23]) of the resource representations of a given URI they have recorded. Moreover, they can resolve a specific version (i.e., a memento) of the resource close to a given time in the past (using a TimeGate resource) [24-25]. An archival replay system, when replaying a memento (an archived version of a representation of a resource), usually also reports first, last, previous, and next mementos in a "Link" response header using corresponding link relations. In traditional web archival replay systems, an index is used that is sorted primarily on a key, called, Sort-friendly URI Reordering Transform (SURT) [26], and secondarily on datetime (in `YYYYMMDDhhmmss` format). This ensures that all the observations of the same URI have spatial locality in the index, making it easier to list all the versions or locate a specific version close to a given date and time. Those index entries point to corresponding byte offsets and chunk sizes in corresponding WARC files where the archived data is stored (often in compressed format).

We used a similar idea in our IPWB system during our initial exploration of decentralized web archiving, but we loaded individual records in IPFS and replaced WARC references in the index file with corresponding IPFS CIDs instead. The downside of this approach was centralization of the index. We are changing it in this work by making IPAROs hold necessary references to prior versions for traversal as shown in Fig. 2.

By leveraging the "Link" header concept of the Memento protocol we can add references in the header part of an IPARO to a number of prior

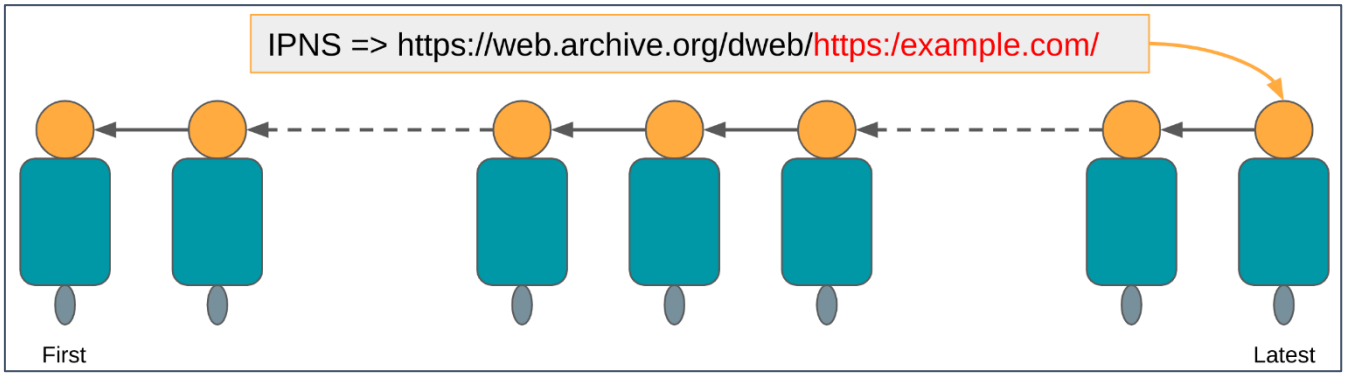


Fig. 2: An Immutable Linked List of IPAROs with an IPNS Entry Pointing to the Head

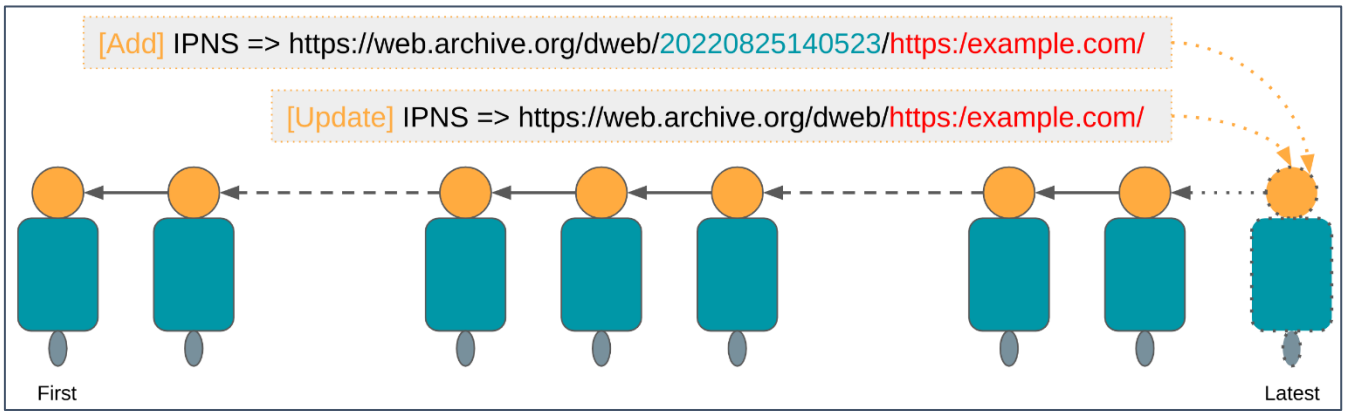


Fig. 3: Adding a New IPARO to an Existing Linked List with an Updated and an Added IPNS Entries

observations of the resource held by that IPARO. These references (i.e., IPFS CIDs) can point to the known immediate previous observation, the very first observation, and number of other random IPAROs in the chain along with their corresponding observation dates and times. Reference to the previous observation of each IPARO would make a singly linked list, allowing linear traversal in the version history back in time as shown in Fig. 2. Adding reference to the first observation is useful because there is usually a significant interest in knowing when a resource was observed the very first time (e.g., to assess the age of a resource). A random set of prior version links are added for iteration efficiency and chain resilience (discussed in sections III-B3 and III-B4).

1) Addition - When a new observation of a resource is recorded, the corresponding new IPARO is appended to the head of the kinked list (if one exists for the given resource identifier). First, an IPNS query is performed to find the CID of the latest IPARO of the given URI (if exists). This CID is then added to the header of the newly created IPARO as the previous link reference (along with additional links like the first one of the list) and then the IPARO is added to IPFS. Finally, the IPNS entry is updated to point to the CID of the newly added IPARO as the latest observation. Moreover, an additional permalink IPNS entry is added with specific date and time to point to the newly added IPARO as shown in Fig. 3.

IPAROs are immutable objects, so adding links to future observations of a resource in prior versions is not possible. This means chain traversal will be unidirectional and easy to report the first and previous versions. However, in the next section we discuss how it is still possible to report the last and next versions of a given IPARO.

2) Routing - Traditional web archives, such as Wayback Machine, support the following templates of archival playback URIs (where URI-R is the original resource URI):

```
<BASEURL>/*/<URI-R>
    [TimeMap/Calendar view]
<BASEURL>/<URI-R>
    [Redirect to the last memento]
<BASEURL>/<DATETIME>/<URI-R>
    [Permalink to the DATETIME version]
```

To support the first of these three representations, we will need the list of CIDs of

IPAROs of all the versions of a given original resource URI and their corresponding archival dates and times. We talk about collecting this data efficiently in the next section.

their collections and redirect to a more appropriate permalink (with a different datetime). In order to support this use case, we will need some traversal of the chain, that we discuss in the next section.

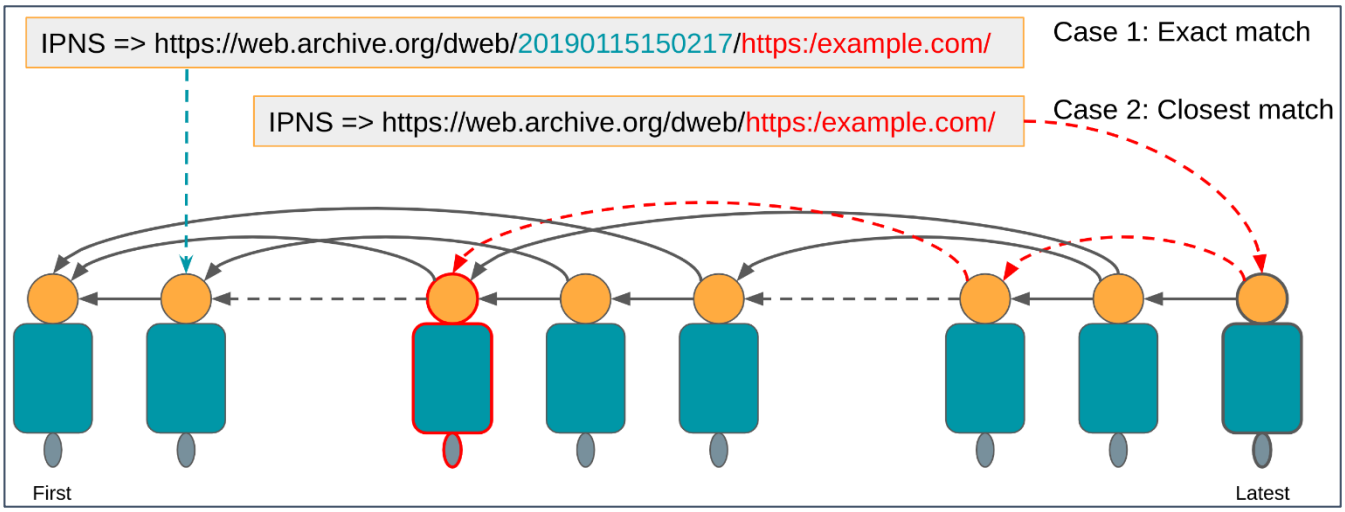


Fig. 4: Illustration of Discovering an Exact or Closest IPARO w.r.t. a Given Date and Time

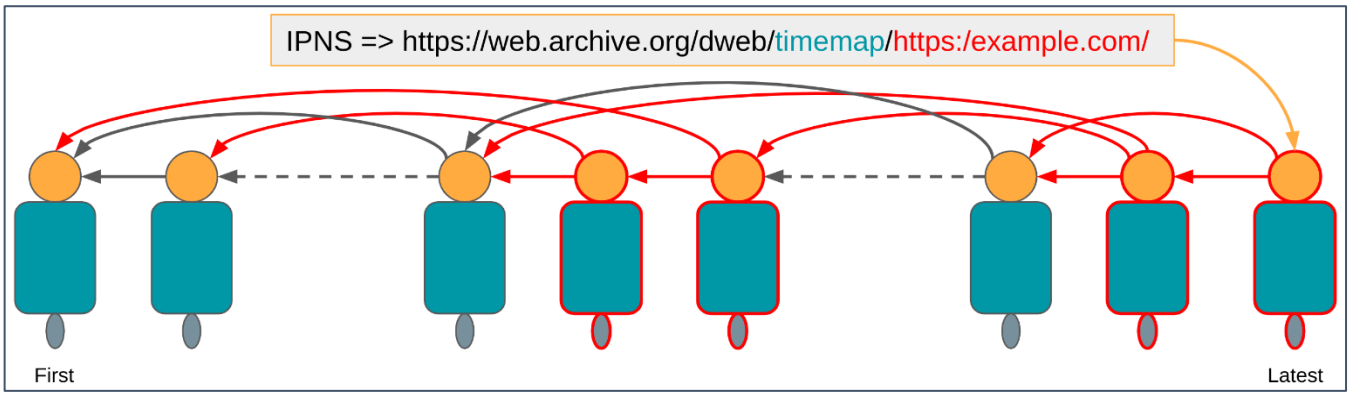


Fig. 5: An Efficient Approach to Construct a Complete TimeMap by Fetching the Linked List Only Partially

The second and third URI templates can be supported using IPNS. When an IPARO is stored in IPFS, an IPNS entry of the third URI form is created that points to the CID of the IPARO. This IPNS entry serves as a permalink to that version. If the newly created IPARO is also the most recent version of the resource, then an IPNS entry of the second URI form is created or updated (if one already exists) to point to it. This way, there is always a quick way to access the most recent version (or the head of the linked list) of every archived resource. Once we have the head of the linked list, we can access all the prior versions of the resource by following the chain via the "Link" header.

If an archived resource is not present at the exact `<DATETIME>` of the third form of the URI, traditional web archives do not treat it as a permalink. Instead, they try to find the closest archive to that datetime in

In traditional web archives it is common to have more than one observation of a given resource with the same datetime, especially when the requests are redirected to a canonical form (e.g., http/https or www/apex origin changes) of the URI by the original host within a second. It is important to maintain the IPNS mapping of the third form of the URI and treat it as immutable. To minimize such collisions, we can use a finer temporal granularity, which is an information that is often recorded, but is not exposed in the URI or index of the traditional web archives for historical reasons (also, HTTP semantics support temporal granularity of one second). If we encounter a collision, despite a finer temporal granularity, we can link IPAROs with "see also" semantics, instead of overwriting the IPNS mapping.

It is worth noting that this decentralized model of archiving democratizes web archiving and makes it

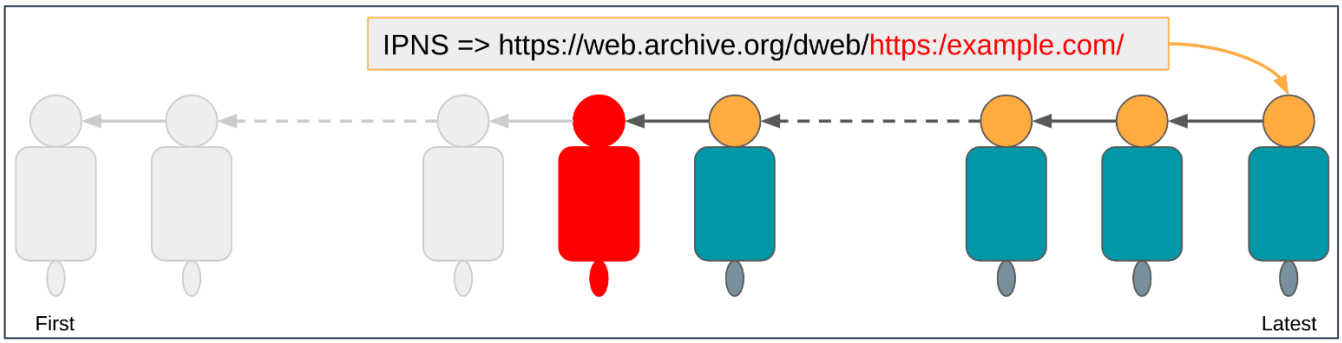


Fig. 6: An Incomplete Linked List Traversal Due to Inaccessible IPAROs with Only the Previous Links

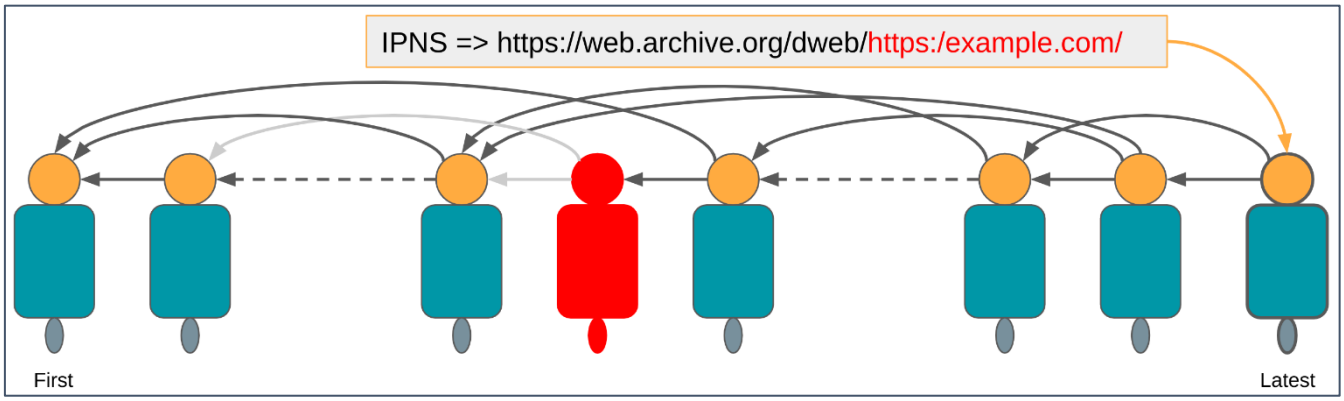


Fig. 7: A Resilient Linked List with Multiple Links to Prior Versions

available as long as operators own a domain name for IPNS record mappings. It allows individual players to have their own chains of observations of the web while allowing memento aggregators [27-28] to mix and match consolidated web archives from many different sources either on-the-fly or by creating IPNS entries under their domains (or even by caching popular resource chains locally).

3) Iteration - So far, we have established that we can have quick access to the most recent archival version of a resource using IPNS and the very first version by accessing any version and reading the appropriate link reference. Moreover, we can also access any IPARO if we know its IPNS permalink as shown in Fig. 4 (Case 1). However, there are times when an application needs a list of all the versions of a resource or discover a version closest to a given date and time. This is possible by linear traversal of the linked list, starting from the most recent version and going backward, one version at a time, using the previous version references. However, this process would be slow for resources with too many archived versions. Also, if any IPARO in the linked list is inaccessible then prior versions will not be reachable.

A more performant approach would be to leverage all the memento link relations stored in IPAROs, not just the one that points to the previous version. For example, to construct the TimeMap (i.e., the list of all or most of the observations), we can: 1) start with the most recent IPARO of the given URI and fetch its headers, 2) add CIDs and corresponding date and times to a set of discovered versions, 3) pick a random CID, that is not yet fetched, and fetch its headers to discover more datetime/CID pairs to be added to the discovery set, and 4) repeat the process until growth of the discovery set stops or slows down significantly. Depending on how densely the "Link" header of IPAROs is populated with prior records, it may require fetching only a fraction of the chain to know all the datetime/CIDs pairs as shown in Fig. 5.

Similarly, to find a IPARO closest to a given date and time: 1) start with the most recent versions and collect all the CID/datetime pairs from its headers, 2) then select the smallest timestamp from the set that is greater than or equal to the desired timestamp and fetch headers of that IPARO, and 3) repeat the process until an exact match is found or all the timestamps discovered from the IPARO are smaller than the desired one. Either that IPARO or the one before it (referenced as previous link) will be the

closest one as shown in Fig. 4 (Case 2). This is a greedy algorithm for the task, but there can be more efficient approaches depending on what strategies were used to populate the "Link" header of IPAROs at the time of their creation.

4) Resilience - It is not guaranteed in a peer-to-peer network that all the nodes containing historical IPAROs of a resource be available and accessible all the time. If the data is not replicated on enough nodes, it is possible that some IPAROs may not be reachable when iterating over the linked list backward using the previous version references as shown in Fig. 6. This is where a strategically selected set of references to various prior versions stored in every IPARO can play a critical role in avoiding roadblocks or broken chains. Those additional links will allow jumping past the missing link of the chain and continue the iteration as shown in Fig. 7, resulting in the loss of just a few IPAROs.

There is an inherent trade-off between storage overhead vs. resilience of the chain and speedy reconstruction of TimeMaps. On the one extreme, each IPARO may only store the reference to its predecessor to have small overhead but has the risk of broken chains. On the other extreme, each IPARO may store references to all the IPAROs prior to it for a given resource, making it extremely resilient, but having an $O(N^2)$ storage overhead. An optimal configuration can likely be achieved with constant (or amortized constant) reference storage overhead with strategic selection of candidates. Various policies should be evaluated in the future to see which ones work well.

C. *Namespacing*

Namespacing is an alternate approach of web archiving and replay without an explicit local index. It is orthogonal to the approach established with the linked list and IPNS. This approach is designed to work on small collections, stored in local IPFS nodes (unless there is an API to perform query for CID prefixes in the peer-to-peer network).

1) Media Type Namespace - In section III-A we described the generic concept of IPMTs and discussed a few media types specific to web archiving. An IPFS node may store all sorts of data. In order to group the data of our interest, we can attach a namespace to them. In this case, we can assign a specific bit sequence at a specific part of the CID (e.g., first six bits be "*010011*") to be associated with certain media types of our interest as shown in Fig. 8. When creating an IPARO we add a nonce at the end of it and keep changing the value of the nonce until the generated CID of the IPARO meets the archival

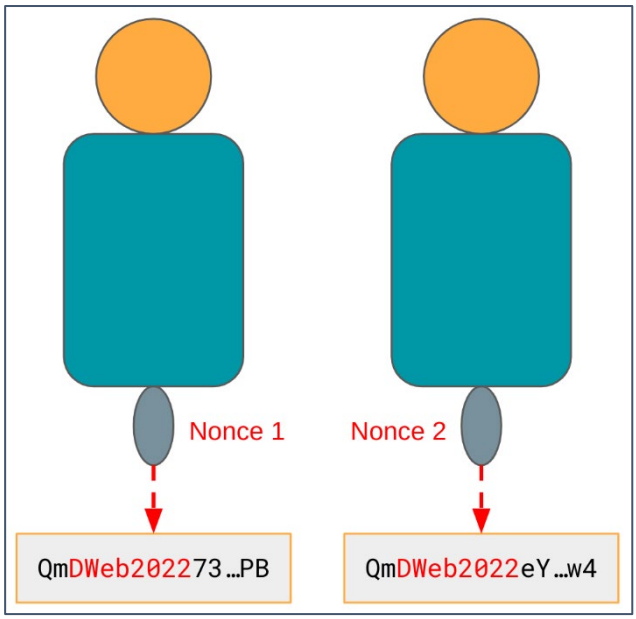


Fig. 8: Two IPARO CIDs With a Desired Common Substring

namespace bitmap. We add nonce at the end for efficient CID calculation, adding it in the header section of an IPARO would require hashing everything each time a nonce is changed. Similarly, for other applications and media types some other namespaces can be assigned. Now, when a web archiving system interacts with the IPFS store, its sample space is smaller.

2) URI Sub-Namespace - To make the lookup space even smaller, we can further assign a few bits as a sub-namespace for original resource URIs. The resource URI is first canonicalized, then hashed, and finally the first few bits of the hash are selected to be used as the sub-namespace (e.g., a URI's hash might have the first ten bits as "*1110010110*"). In this case, the IPARO will be stored with a nonce value that causes the CID to have 16 bits of prefix of "*0100111110010110*".

At the time of replay, a similar calculation can be performed to identify the candidate namespace then the system can query for all the CIDs of the node that have the desired bitmap match. This will likely be a small set of nodes, which can then be inspected on-the-fly to discard any irrelevant IPAROs.

3) Collisions - Such bitmap-based namespacing is prone to collisions, especially, when the namespace is short and the IPFS node contains too many objects. A larger namespace will minimize the probability of collisions but will come with the added compute cost of finding the right nonce value at the time of IPARO creation. This compute cost would double, and the probability of collisions would halve with each bit of increment in the namespace. Any remaining collisions must be identified and discarded on the fly at the playback time.

D. Composition and Decomposition

So far, we have treated IPAROs as a blob containing some headers, the actual IPMT payload (described in section III-A), and some trailers like nonce (as described in section III-C). However, it might be more space-effective to split the blob in strategically identified chunks and store them in pieces in the IPFS store. At the time of playback, those pieces need to be put together to form the complete IPARO.

1) Deduplication - In our previous exploration, in IPWB, we split HTTP response records in two parts, headers and payload, and stored them separately. Consequently, our index had two CIDs for each observation. We did it because frequent observations of the same resource might result in the same payload or even the same payload can be seen across resources (such as soft-404 responses [29]). However, each HTTP response contains a "Date" header which indicates the time when the response was generated, which often changes every second. By splitting headers from payloads, we were able to achieve deduplication on payload while storing unique headers each time.

In this work we propose a flexible approach of chunking IPAROs. At the time of storing an IPARO, one may choose to store it as a single object or split it in as many chunks as desired. For example, one may choose to split a Memento IPARO in pieces like the IPMT header, request headers, any request payload, selected response headers that are often unchanged, remaining changing headers, response payload, and trailers as shown in Fig. 9. If the application identifies parts of the payload that are common across other responses, it can split the payload on those strategic points as well to achieve increased deduplication as the collection grows.

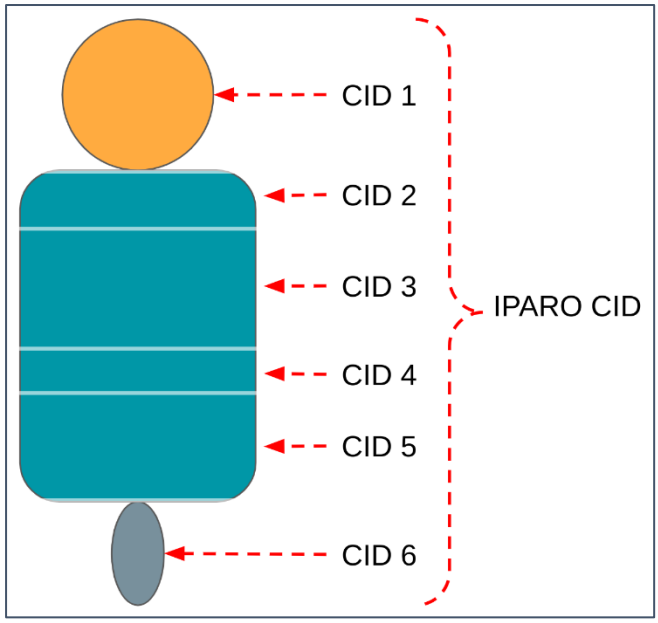


Fig. 9: Strategic Blocks for Efficient Merkle DAG Deduplication

2) Substitution by Reference - In order to put the pieces of an IPARO together (if it is stored as chunks), one possibility is to use a templating language and replace extracted pieces in the container object with their corresponding CIDs. At the time of replay, first fetch the container object, process it to find any inline references, fetch those pieces, and replace them in-place to form the original complete IPARO.

This approach would require a rehydration step to process each IPARO and populate any inline references. Using a templating language will come with the additional overhead of placeholder markers, any escape sequences, and the space needed for CIDs.

3) Concatenation - This approach leverages a feature of IPFS in which it is possible to tell the system to create blocks of desired sizes instead of using the default block size to construct the Merkle DAG. To store data with strategically designated block sizes, the pieces of an IPARO are determined as described in section III-D1, then those pieces are stored one by one as independent objects as shown in Fig. 9 (CID 1-6). In the next step the whole IPARO is stored, but this time the system is told to have blocks of provided sizes (that align with the split points). This way, the system realizes that all the blocks already existed in the Merkle DAG, so no more data is stored, but a new CID is returned that represents the complete IPARO as shown in Fig. 9 (IPARO CID) [30]. At the time of replay, no additional post-processing or rehydration is needed as fetching the composite

CID will return concatenated data as if it were a single piece of data in the IPFS store.

IV. Conclusion and Future Work

We have proposed a decentralized version tracking system using the existing primitives of IPFS and IPNS. While our description talks primarily about archived web pages, we have proposed the concept of IPMT and namespacing so that it can be used in other applications that require versioning, such as a wiki or a collaborative code tracking system. Our proposed system does not rely on any centralized server for archiving or replay of the content. The system continues to allow aggregators to play their role from which both large and small archives can benefit and flourish. Future plans include a proof-of-concept implementation and evaluations of various aspects.

Acknowledgments

We thank Juan Benet from Protocol Labs, Dr. Mat Kelly from Drexel University, and Dr. Michael Nelson from Old Dominion University for their initial reflections on the idea and reviews. We thank Ilya Kreymer from Webrecorder Software for timely mention of the possibility of storing data in IPFS with custom block sizes. Also, we thank Irakli Gozalishvili from Protocol Labs for the reference to the Bamboo data structure. This work is supported in part by Protocol Labs and Filecoin Foundation, we thank Dietrich Ayala for the coordination.

References

[1] J. Benet, "IPFS - Content Addressed, Version, P2P File System," Tech. Rep. arXiv:1407.3561, 2014.

[2] S. Alam, M. Kelly, and M. L. Nelson, "InterPlanetary Wayback: The Permanent Web Archive," in Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries, ser. JCDL '16, 2016, pp. 273–274.

[3] M. Kelly, S. Alam, M. L. Nelson, and M. C. Weigle, "InterPlanetary Wayback: Peer-to-Peer Permanence of Web Archives," in Proceedings of the 20th International Conference on Theory and Practice of Digital Libraries, 2016, pp. 411–416.

[4] S. Alam, M. Kelly, M. C. Weigle, and M. L. Nelson, "InterPlanetary Wayback: A Distributed and Persistent Archival Replay System Using IPFS," DWeb Summit '18, 2018.

[5] S. Alam, M. Kelly, M. C. Weigle, and M. L. Nelson, "InterPlanetary Wayback: The Next Step Towards Decentralized Web Archiving," IPFS Lab Day '18, 2018.

[6] S. Alam, "IPNS Blockchain," https://github.com/oduwsdl/IPNSBlockchain, 2018.

[7] S. Alam, "Implementing History Aware IPNS Via Certificate Transparency Log Data Structure Trillian," https://discuss.ipfs.tech/t/implementing-history-aware-ipnsvia-certificate-transparency-log-data-structure-trillian/5756, 2019.

[8] S. Alam, M. Kelly, M. C. Weigle, and M. L. Nelson, "Client-Side Reconstruction of Composite Mementos Using ServiceWorker," in Proceedings of the 17th ACM/IEEE-CS Joint Conference on Digital Libraries, ser. JCDL '17, 2017, pp. 237–240.

[9] M. Kelly, M. L. Nelson, and M. C. Weigle, "Making Enterprise-Level Archive Tools Accessible for Personal Web Archiving," https://www.slideshare.net/matkelly01/makingenterpriselevel-archive-tools-accessible-for-personal-webarchiving, 2013.

[10] J. A. Berlin, M. Kelly, M. L. Nelson, and M. C. Weigle, "WAIL: Collection-Based Personal Web Archiving," in Proceedings of the IEEE/ACM Joint Conference on Digital Libraries (JCDL), 2017, pp. 340–341.

[11] I. Kreymer, "Webrecorder Tools," https://webrecorder.net/tools, 2020.

[12] J. M. Patel, "Introduction to Common Crawl Datasets," pp. 277– 324, 2020.

[13] M. Phillips and S. Alam, "Hosting the End of Term Web Archive Data in the Cloud," https://netpreserve.org/ga2022/wac/abstracts/#Session_9, 2022.

[14] S. Alam, M. L. Nelson, H. Van de Sompel, L. L. Balakireva, H. Shankar, and D. S. H. Rosenthal, "Web Archive Profiling Through CDX Summarization," International Journal on Digital Libraries, vol. 17, no. 3, pp. 223–238, 2016.

[15] S. Alam, M. C. Weigle, M. L. Nelson, F. Melo, D. Bicho, and D. Gomes, "MementoMap Framework for Flexible and Adaptive Web Archive Profiling," in Proceedings of the 19th ACM/IEEECS Joint Conference on Digital Libraries, ser. JCDL '19, 2019, pp. 172–181.

[16] ISO 28500:2017, "WARC File Format," https://iso.org/standard/68004.html, 2017.

[17] A. Meyer, "Bamboo," https://github.com/AljoschaMeyer/bamboo, 2019.
[18] IPLD, "InterPlanetary Linked Data (IPLD)," https://ipld.io/docs/, 2021.
[19] Internet Archive, "CDX File Format," http://archive.org/web/researcher/cdx_file_format.php, 2003.
[20] I. Kreymer and E. Summers, "Web Archive Collection Zipped (WACZ)," https://specs.webrecorder.net/wacz/1.1.1/, 2021.
[21] J. Odvarko, A. Jain, and A. Davies, "HTTP Archive (HAR) Format," https://w3c.github.io/webperformance/specs/HAR/Overview.html, 2012.
[22] R. Sanderson, P. Ciccarese, and B. Young, "Web Annotation Data Model," https://www.w3.org/TR/annotation-model/, 2017.
[23] H. Van de Sompel, M. L. Nelson, and R. Sanderson, "HTTP Framework for Time-Based Access to Resource States – Memento," RFC 7089, Internet Engineering Task Force, 2013.
[24] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. L. Balakireva, S. Ainsworth, and H. Shankar, "Memento: Time Travel for the Web," Tech. Rep. arXiv:0911.1112, 2009. [Online]. Available: https://arxiv.org/abs/0911.1112
[25] M. L. Nelson and H. Van de Sompel, "Adding the Dimension of Time to HTTP," in The SAGE Handbook of Web History, 2018.
[26] K. Sigurðsson, M. Stack, and I. Ranitovic, "Heritrix User Manual: Sort-friendly URI Reordering Transform," http://crawler.archive.org/articles/user_manual/glossary.html#surt, 2006.
[27] S. Alam and M. L. Nelson, "MemGator - A Portable Concurrent Memento Aggregator: Cross-Platform CLI and Server Binaries in Go," in Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries, ser. JCDL '16, 2016, pp. 243–244.
[28] R. Sanderson, H. Van de Sompel, and M. L. Nelson, "IIPC Memento Aggregator Experiment," http://www.netpreserve.org/sites/default/files/resources/Sanderson.pdf, 2012.
[29] L. Meneses, R. Furuta, and F. Shipman, "Identifying 'Soft 404' Error Pages: Analyzing the Lexical Signatures of Documents in Distributed Collections," in Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries, ser. TPDL '12, vol. 7489, 2012, pp. 197–208.
[30] I. Kreymer, "IPFS Composite File Utilities," https://github.com/webrecorder/ipfs-composite-files, 2022.
[31] S. Alam, M. C. Weigle, M. L. Nelson, M. Klein, and H. Van de Sompel, "Supporting Web Archiving via Web Packaging," Exploring Synergy between Content Aggregation and the Publisher Ecosystem (ESCAPE) Workshop, 2019. [Online]. Available: https://arxiv.org/abs/1906.07104.
[32] A. Potsides, M. Rataj, S. Loeppky, D. Norman, and E. Lee, "State of IPFS in JS," https://blog.ipfs.tech/state-of-ipfs-in-js/, 2022.

CONTENT-BASED CHARACTERIZATION OF THE END OF TERM WEB ARCHIVE

Mark E. Phillips
University of North Texas
USA
mark.phillips@unt.edu
0000-0002-9679-6730

Kristy K. Phillips
University of North Texas
USA
kristy.phillips@unt.edu
0000-0002-3750-3176

Sawood Alam
Internet Archive
USA
sawood@archive.org
0000-0002-8267-3326

Abstract—Since 2008, the End of Term Web Archive has been gathering snapshots of the federal web, consisting of the publicly accessible .gov and .mil websites. In 2022, the End of Term team began to package these crawls into a public dataset which they released as part of the Amazon Open Data Partnership program. In total, over 460TB of WARC data was moved from local repositories at the Internet Archive and the University of North Texas Libraries. From the original WARC content, derivative datasets were created that address common use cases for web archives. These derivatives include WAT, WET, CDX and a format called a WARC Metadata Sidecar. This WARC Metadata Sidecar includes content-based characterizations of files held in the archive, including character set, language, file format identifier, and soft 404 detection. This paper describes the decisions made in the creation of these derivatives, the technologies used, and introduces the WARC Metadata Sidecar, which presents a useful approach for creating and storing auxiliary metadata for web archives.

Keywords – web archives, End of Term Web Archive, WARC Metadata Sidecar

I. INTRODUCTION

The End of Term (EOT) Web Archive is a collection of web crawls of all publicly available federal websites on the .gov and .mil domains collected concurrently with each presidential election since 2008. This project to document the United States federal web is the result of a collaboration between the Internet Archive, the Library of Congress, the University of North Texas, and many other organizations. The archive includes four web crawls, three of which were collected during years in which a new president was elected (2008, 2016, 2020), and one that was collected in a year in which the current president was re-elected (2012). During the years in which a new president is elected, this archive serves to document the effect of the transition on public websites. When an incumbent president is re-elected, the web crawl documents any changes made to the federal web over the four years since the previous election.

In 2022, the Internet Archive and the University of North Texas began working to create the End of Term Web Archive Dataset, a more accessible dataset of the content found in the EOT Web Archive. This dataset overcomes the logistical challenges faced by users interested in using the archive for computationally-focused research and allows open access to a large, longitudinal dataset of the federal web.

The full dataset is available with a Creative Commons CC0 1.0 Universal (CC0 1.0)[1] Public Domain Dedication and is downloadable from the End of Term Website[2]. A record for the dataset is also available in the Registry of Open Data on AWS[3].

II. RELATED WORK

The idea of a metadata sidecar file is not new. Referred to as a sidecar, buddy, or connected files, they allow for additional metadata to be stored alongside the primary file in situations where either the primary file does not include a method for storing arbitrary metadata, or in situations where you do not want to change the original files.

Perhaps the most common sidecar file is part of the suite of specifications that formalize file formats

[1] Creative Commons CC0 1.0 Universal https://creativecommons.org/publicdomain/zero/1.0/

[2] End of Term: Data https://eotarchive.org/data/

[3] Registry of Open Data on AWS https://registry.opendata.aws/eot-web-archive/

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

in Adobe's Extensible Metadata Platform (XMP) [1]. Generally, these files have the extension .xmp, and are stored in the same directory as the file that they reference. An XMP sidecar file is an XML file that stores information about the original file or change instructions from non-destructive editing tools like Adobe Bridge, Adobe Lightroom, or other tools.

Within the web archiving space, several other derivative sidecar files are commonly produced that either provide easier access to data within the original Web ARChive file format (WARC), or include a processed dataset generated from those WARC files. For example, the most common derivative file generated from the WARC records is a CDX file. A CDX file, which is a column-based text file that is used to create an index of the contents of WARC files, facilitates lookup and replay of archived web resources. Two other derivative sidecar formats common in web archiving are the Web Archive Transformation (WAT) file and the Web Archive Extracted Text (WET) file. These derivative files are often named in such a way that it is clear to users which WARC files they were derived from. For example, WAT file names typically take the base WARC name and add *.wat.gz*, the WET file names add *.wet.gz*, and the CDX file names add *.cdx.gz*. These files are typically compressed with GZip, though by different means. WAT and WET files follow the same practice as WARC records and use a record-at-time compression, while the CDX files use a full file compression. Though these filename patterns are not mandatory, they are standard practice in the web archiving community, with several software packages writing these by default (hadoop-tools, cdx-indexer, others).

It is common practice to generate derivative files for web archives, in part to improve access to the underlying data stored in the primary WARC files. This is done for several reasons, the foremost being that WARC files in web archives generally require large amounts of storage that may be beyond what a researcher interested in working with the archive might have available. To cut down on file size, derivatives that only contain a portion of the dataset are generated. For example, in a WAT file, the links, link text, and HTML metadata is the content primarily extracted. This usually results in a significant decrease in the amount of storage space required, as the WAT file only contains data extracted from certain formats like HTML, while large binary files like PDF, JPEG, or MP4 files are not included. Similarly, the WET file only contains text extracted from HTML and TXT files, so the resulting derivative file is much smaller than the original WARC file. An example of the size difference can be seen in the End of Term Dataset, where a WARC file[4] from the EOT-2020 crawl has a size of 953.7 MiB, and its corresponding WAT, WET, WARC Metadata Sidecar, and CDX files have sizes of 449.5 MiB, 82.7 MiB, 40.5 MiB, and 3.9 MiB respectively.

Over the past decade, the Archives Unleashed Project [2] has developed a toolkit and services for generating and using derivative files from web archives. This project has worked to improve the capacity for researchers to use web archives in a wide variety of research areas. The ability to work with extracted derivatives generally covers a wide range of use cases and can be a great way to encourage research interest in web archives as a data source.

The Library of Congress Web Archive and the UK Web Archive (UKWA) are among a growing group of national web archiving programs that are generating sample datasets and derivatives of their collections for use by researchers and scholars [3], [4]. In some cases, institutions are not able to directly share their web archives due to copyright or other rights restrictions. These restrictions require different approaches to data sharing. One of these approaches is to share derived metadata from the source material, which enables non-consumptive use of the underlying resources. These derivatives can also help overcome challenges researchers face in working with these web archives due to their size and scale.

Perhaps the best example of an organization that provides ample derivative formats for web archives is the Common Crawl initiative. Common Crawl operates monthly web-scale crawls of primarily text-based content like HTML, TXT, and PDF files, then makes these crawls publicly available. In addition to the WARC content, they generate WAT, WET, CDX, and Parquet files. Parquet files provide an index of the content in a WARC file using a column-oriented storage structure. In addition to these standard

[4] EOT20-20201009-crawl800_EOT20-20201009165718-00000.warc.gz

derivatives, Common Crawl provides content-based characterizations of the files they harvest at crawl time. For example, for each HTML and TXT file that they harvest, they perform content-based language identification, character set detection, and MIME type detection [5]. Because this characterization is done at crawl time, Common Crawl can store these additional metadata fields as WARC '*metadata*' records inside the primary WARC file without having to store them as a sidecar file. Once extracted, Common Crawl makes these additional metadata fields available in the CDX and Parquet indexes.

III. Overview of the EOT Web Archive Dataset

The End of Term Web Archive Dataset contains the 2008, 2012, 2016, and 2020 web crawls that make up the End of Term Web Archive. These have been collocated in the Amazon cloud as part of Amazon's Open Data Program [6]. The EOT Dataset is available using standard HTTP or an S3 client for download. The dataset is grouped so that a user can decide how much data they want to download, from the entire dataset or the data from a given election cycle, to a dataset collected by a specific crawling partner within an election cycle. The primary dataset contains ARC/WARC files, with one derivative in the format WAT, WET, CDX, and WARC Metadata Sidecar (META) created for each primary file. While the formats used in the dataset described in this paper are common in the web archiving community, it is useful to introduce the formats for those interested in the dataset that are not as familiar with the formats. Additionally, the documentation for these formats can be too dense to serve as a brief introduction to them. The sections below give a brief overview of these different files and derivatives.

A. WARC

The bulk of the dataset is housed in the WARC format. This is the standard format used in the field of web archiving to store harvested data and was designed specifically for this purpose. It contains individual WARC records that are compressed with GZip and concatenated into a single WARC file. There are different WARC record types, including '*warcinfo*,' '*response*,' '*resource*,' '*request*,' '*metadata*,' '*revisit*,' '*conversion*,' or '*continuation*'. Many tools are available for reading and writing the WARC format. The WARC format is an ISO Standard (ISO 28500:2017) and is maintained by the International Internet Preservation Consortium [7].

The WARC format was standardized in 2009 and because of this, the web crawl from 2008 contains ARC files in addition to WARC files. ARC is the predecessor to the WARC format and many web archiving institutions have chosen to maintain these original formats instead of migrating them to the WARC format. The ARC format is typically supported by tools written for the WARC format because they are so similar. The EOT Dataset maintains the original file formats and does not include any format migration from ARC to WARC in cases where ARC files were created during the initial web crawl.

B. WAT

The Web Archive Transformation (WAT) derivative is generated for each of the primary files in the dataset. These files align with the primary WARC files and provide extracted metadata and link structures from HTML content. These extractions can be used for various activities where the full text of the resource is not needed but the links from that resource and their accompanying anchor text is desired. For example, WAT files are useful for building link graphs [8]. The WAT file is generally a fraction of the size of the original WARC file. To keep alignment with the original files, the *.warc.gz* from the primary WARC is changed to *.warc.wat.gz* for the WAT file.

C. WET

The Web Extracted Text (WET) derivative is extracted text content from HTML and TXT formats in the primary WARC files. This extracted text is useful in many situations where the full structure of the HTML resource is not required. They are also streamlined for processing because they do not contain records for non-HTML and non-TXT resources. This makes the overall size of a WET file much smaller than the original WARC and generally smaller than the corresponding WAT derivative. To keep alignment with the original files, the *.warc.gz* from the primary WARC is changed to *.warc.wet.gz* for the WET file.

D. CDX

A CDX file is created that contains a row-oriented index of the WARC records inside of the WARC file. Each row contains multiple pieces of information

related to the harvested content. These include: the URL, a reversed and sort friendly URL format called a Sort-friendly URI Reordering Transform (SURT), a datetime, HTTP response code, and MIME type supplied by the server the resource was harvested from, the number of bytes in the WARC record's content payload, the offset in the WARC record, the payload digest, and the WARC file path. These are generally sorted using the SURT URL key and datetime columns and then further grouped together to create indexes that can drive replay systems such as Open Wayback [9] or pywb [10]. There are several common row configurations of a CDX file, including nine-field, eleven-field, and the CDXJ configuration, which allows for more arbitrary metadata to be stored beyond the typical nine or eleven fields. The layout of a CDXJ row is the sortable URL, a timestamp, and then a single-line JSON object containing additional standard metadata fields used for replay and additional fields as needed. To keep alignment with the original files, the *.warc.gz* from the primary WARC is changed to *.cdxj.gz* for the CDX file.

E. Parquet Index

The Parquet format is a column-oriented data file format designed for efficient data storage and retrieval [11]. It is used to provide a different way of accessing the data held in the CDX derivatives. The Parquet format is used in many big-data applications and is supported by a wide range of tools. This derivative allows for arbitrary querying of the dataset using standard query formats like SQL and can be helpful for users who want to better understand what content is in the EOT Dataset using tools and query languages they are familiar with.

F. WARC Metadata Sidecar

The WARC Metadata Sidecar, referred to as the META derivative, contains content-based characterizations of the WARC records. It is described in detail in the following section.

IV. WARC METADATA SIDECAR (META)

The WARC Metadata Sidecar, referred to as the META format in the EOT Dataset, was created as a way for the team to address the problem of generating and storing additional metadata for WARC Records from the primary WARC files. As explained in the background section, the concept of a metadata sidecar file is not a new idea but an implementation of an existing concept. Other derivatives like WAT and WET essentially serve a similar function to a WARC Metadata Sidecar file, though they don't contain the same information. A WARC Metadata Sidecar file contains content-based characterizations generated using tools applied to the data, rather than a simple distillation of data from the original resource, as is found in WAT and WET files.

The metadata sidecar files contain content-based characterizations of the *response* and *resource* WARC record types. The resulting metadata fields are stored in a '*metadata*' WARC record using the warc-fields format which is a key/value format used within the WARC records themselves.

For the creation of the EOT Dataset, an open-source tool written in Python called *warc-metadata-sidecar.py* [21] was created that processes a single WARC file and generates a corresponding WARC Metadata Sidecar. The resulting filename changes the *.warc.gz* extension to *.warc.meta.gz*, which keeps the new file aligned with the original in a similar way as is done with WAT, WET, and CDX files for the dataset.

Because the project required the creation of several hundred thousand WARC Metadata Sidecar files, the team made use of mrjob [12], a Python framework for writing and running distributed computing jobs using Apache Hadoop [13]. The team used a small, 5-node Hadoop cluster housed at the UNT Libraries for the processing of all the primary WARC files.

A. Character Sets

Character set detection is implemented with the Python library *Chardet: The Universal Character Encoding Detector* [14]. This library is a continuation of the work by Mark Pilgrim and his original port of the C++ universal character encoding detector from Mozilla that he called *chardet* [15]. The output of this library is a prediction of the most likely character set of the input text and the confidence that the tool has in its prediction. These two values are stored under the key Charset-Detected in the payload of the WARC metadata record.

B. Language Identification

Language identification is accomplished using the Python bindings for the Compact Language

Detector 2 (CLD2) library [16]. CLD2 can detect over 80 languages in Unicode UTF-8 text and can work with either HTML or XML. It makes use of a Naive Bayes classifier and different token algorithms. This tool was originally introduced by Google as part of the Chrome browser where it is used for language detection in that application. An updated method of language identification has been introduced called CLD3, which makes use of neural networks instead of Bayesian classifiers for language prediction. The EOT team chose to work with the CLD2 implementation because it had existing functionality for working with HTML and XML formats, while CLD3 requires conversions from those formats to UTF-8 to be done outside of the library. The output of the library is a list of the predicted languages, a score for that language, if the prediction should be considered reliable, and how much of the input text is represented by that language. The metadata sidecar takes the top three languages for a resource and stores those under the key Languages-cld2 in the payload of the WARC metadata record.

C. File Format Identification

File format identification is accomplished using the tool *Format Identification for Digital Objects* (fido) [17]. This is a tool originally developed by Adam Farquhar of the British Library and now maintained by the Open Preservation Foundation. It uses signatures from the PRONOM format registry maintained by the National Archives of the UK [18]. Fido was chosen over similar tools like DROID or Siegfried because it is written in Python and would integrate easily with the other libraries used in the *warc-metadata-sidecar.py* tool. The result of this format identification library is both a MIME Type for the format and the unique PRONOM identifier. The metadata sidecar stores the PRONOM identifier under the key Preservation-Identifier and stores the MIME Type under the key Identified-Payload-Type, with a label indicating it comes from fido.

D. MIME Type

In addition to the MIME Type that is identified using fido as described above, another MIME Type detection tool is used to provide an additional data point about the MIME type. In this case the python-magic [19] library, which is a Python interface to the libmagic file type identification library, is used. The output of this tool is often at a more general level than the output of fido. For example, python-magic might identify a file as the type *text/html*, where fido might specify the format as being *application/xhtml+xml*. Both outputs are retained for instances where either the specific or more general identification is desired. The MIME type under the key Identified-Payload-Type includes a label to indicate it comes from either fido or python-magic.

E. Soft 404 Detection

Finally, to experiment with identification of the soft 404 phenomenon, this project used the Python tool Soft-404 [20]. The soft 404 phenomenon occurs when a web server responds with an HTTP response code of *200 OK*, but returns a page that indicates that the content is not available instead of returning a *404 Not Found* response code. The Soft-404 library uses a classifier that was trained on 198,801 pages from 35,995 domains, with a 404 page ratio of about 1/3. The EOT Dataset used the provided model for soft 404 detection. The result is a value between 0 and 1 that shows how likely the page is a soft 404 example with scores closer to zero being unlikely and those closer to 1 more likely to be a soft 404. This value is stored under a key of Soft-404-Detected in the payload of the WARC metadata record.

V. Discussion

There are several reasons it is a good idea to do content-based identification. One example is MIME Type identification using actual content over provided values. In this situation, a server can provide a MIME type like application/pdf and a URL such as https://example.com/sample.pdf, but because of an error in confirmation in the web server, or an error in the coding to dynamically generate content, an HTML file reporting the error or unavailability of the page (404) might be returned without reporting the correct MIME Type. Content-based identification in this case can accurately identify the actual MIME type of the content as HTML. This identification is also important for recognizing Soft 404 which can often return a 200 response with a given MIME type but, the content is text/html.

The metadata extracted from WARC records and included in the WARC Metadata Sidecar files can be used to build indexes of the content available in the End of Term Web Archive. As an example of this, we are using the content-based characterization data in combination with the standard data found in a CDX index to build a Parquet index for each of the End of

Term crawls. These Parquet indexes allow users to answer questions related to the web archives that previously would have been challenging to ask, such as ”what MIME types are misreported the most?”, ”what domains have the most misreported content?”, ”what is the prevalence of non-English content in the archive?”, ”what domains have the most non-English content?”, ”which non-English languages are most represented in the archive?”, ”how prevalent are Soft 404’s and which domains have the most instances of them?”, and ”what are the file types present based on file identifiers?” This list can easily go on, and these are questions that can be answered by writing SQL queries to interact with the Parquet index and do not require traversing the dataset as it would have in the past.

VI. Future Work

The WARC Metadata Sidecar file introduces a method for storing metadata from different content-based characterization tools and associating that metadata with the original WARC files that make up a web archive. They provide a logical alignment with WARC records and allow for content-based characterization of content in ways that were previously unavailable, or with new approaches or tools. The new metadata that is generated can be incorporated into indexes that provide opportunities to answer research questions related to large-scale web archives that could otherwise be challenging to answer.

The implementation of WARC Metadata Sidecar files in this project might be improved in several ways. First, the *warc-metadata-sidecar* tool [21], written in Python and then integrated as mrjob jobs on a Hadoop cluster, was successful at processing content at scale. Inefficiencies were recognized as more content was processed, though there are still situations where the tools might need further optimization to deal with the number of files that require characterization. Limitations exist thanks to file formats that are not compatible with the tools used for content-based language identification, like PDF and JPEG files. One way to improve the implementation of the warc-metadata-sidecar.py in the future might be to incorporate tools like the Tika library [22] to convert additional formats like PDF or Microsoft Word Documents into text that can be further characterized. The introduction of a Java-based tool to the process might warrant a change in underlying programming language used for the overall script. Another option is to investigate Python-based converters that can extract text from various binary files so that they can be incorporated into the output.

Future work for this dataset includes generation of host-level and domain-level network graphs that will show the relationships between domains within the EOT Web Archive. This work is expected to continue to leverage existing tools and processes developed by Common Crawl for graph generation. With the complete dataset available in CDX format, overviews of each of the EOT crawl years using CDX summarization tools [23] can be generated. These can be helpful in communicating the content of this dataset to others.

VII. Acknowledgements

This effort would not have been possible without storage support from the Amazon Open Data Program, which provided S3 storage for this initiative. Likewise, this project leaned heavily upon the prior work of the Common Crawl team and adopted their organizational structures, tools, and documentation in building this dataset and providing access to it.

1. REFERENCES

[1] Adobe, “Adobe – XMP developer center.” https://www.adobe.com/devnet/xmp.html, 2023.
[2] Archives Unleashed, “The Archives Unleashed Project.” https://archivesunleashed.org/, 2022.
[3] Library of Congress, “Web archive datasets,” 2023.
[4] UK Web Archive, “UK web archive open data,” 2023.
[5] Common Crawl, “August crawl archive introduces language annotations.” https://commoncrawl.org/2018/08/august-2018-crawl-archive-now-available/, 2018.
[6] Amazon.com, Inc., “Open Data Sponsorship Program,” 2022.
[7] International Internet Preservation Consortium, “The WARC format 1.0.” https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0/, 2022.
[8] Common Crawl, “Common Crawl’s first in-house web graph,” 2017.
[9] International Internet Preservation Consortium, “OpenWayback.” https://github.com/iipc/openwayback, 2021.
[10] Webrecorder, “pywb.” https://github.com/webrecorder/pywb, 2023.
[11] The Apache Software Foundation, “Apache Parquet,” 2023.
[12] Yelp,“mrjob: the Python mapreduce library.” https://github.com/Yelp/mrjob, 2020.
[13] The Apache Software Foundation, “Apache Hadoop,” 2023.
[14] D. Blanchard, “Chardet: The universal character encoding detector.” https://github.com/chardet/chardet, 2022.

[15] M. Pilgrim, Case Study: Porting chardet to Python 3, pp. 253–277. Berkeley, CA: Apress, 2009.
[16] R. Alrfou, “PYCLD2 – Python bindings to CLD2.” https://github.com/aboSamoor/pycld2, 2022.
[17] Open Preservation Foundation, “Format identification for digital objects (fido).” https://github.com/openpreserve/fido, 2022.
[18] The National Archives, “Pronom.” https://www.nationalarchives.gov.uk/PRONOM 2006.
[19] Hupp, “python-magic.” https://github.com/ahupp/python-magic, 2022.
[20] TeamHG-Memex, “soft404: a classifier for detecting soft 404 pages.” https://github.com/TeamHG-Memex/soft404, 2017.
[21] University of North Texas Libraries, “WARC Metadata Sidecar.” https://github.com/unt-libraries/warc-metadata-sidecar, 2023.
[22] The Apache Software Foundation, “Apache Tika.” https://tika.apache.org/, 2023.
[23] S. Alam and M. Graham, “CDX summary: Web archival collection insights,” vol. 13541 of Lecture Notes in Computer Science, pp. 297–305, Springer, 2022.

NOT WELL-FORMED OR INVALID. NOW WHAT?

Towards a formalized workflow for format validation error treatment

Micky Lindlar
TIB Leibniz Information Centre for Science and Technology
Germany
michelle.lindlar@tib.eu
0000-0003-3709-5608

Abstract – File format validation – we all use it and we all run into problems when files do not validate. Though a core process within digital preservation practice, little progress has been made in shared documentation and discussion of processes used to treat file format validation errors. This paper aims to close that gap. A basic workflow for handling validation errors is proposed and visualized, and in a second step tested against two TIFF and two PDF validation errors of varying severity. Observations made are fed back into the workflow diagram. The outcome shall provide a first step towards shared digital preservation practice in the currently largely neglected field of method formalization for file format validation error treatment.

Keywords – file format validation; process formalization; error handling

Conference Topics – We're all in this Together; From Theory to Practice

I. INTRODUCTION

(Digital) interpretability, i.e., correct rendering of digital objects, is one of the core tasks of digital preservation. Checking if files open in a reader is one method to check correct rendering, however, this is a time-intensive process and can only be achieved perfectly if we know what the original is supposed to look like. Even in textual objects, malformed formulas or tables might easily be missed during visual inspection [1],[2]. File format identification and file format validation therefore serve as standard processes to check for a file's structural and syntactical intactness and they are embedded processes in all major end-to-end digital preservation systems [3]. It is thus safe to say that most digital preservation practitioners should be familiar with the tasks.

Since file format identification is based on short pattern recognition such as "magic number" or byte sequence checking, it is a good first indicator of a digital object's file format, but it is by no means proof that the file can actually be opened. The most reliable method to ensure the renderability of a digital object is file format validation, which checks the digital object's internal syntax against the rules outlined in the file format standard or description. Since the development of a validator not only depends on the availability of a file format description, but is also significantly more resource-intensive than identifying and capturing a file format signature pattern, validators are currently only available for a handful of file format families. Those file formats that do have validators available are often also found in "recommended" or "preferred file format lists" [4].

Benefits of a formalized methodology for validation troubleshooting are threefold:

(1) "A picture is worth a thousand words" – a workflow graphic can enable more effective communication about specific errors amongst practitioners. Furthermore, it can help those just starting out to understand the process better.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

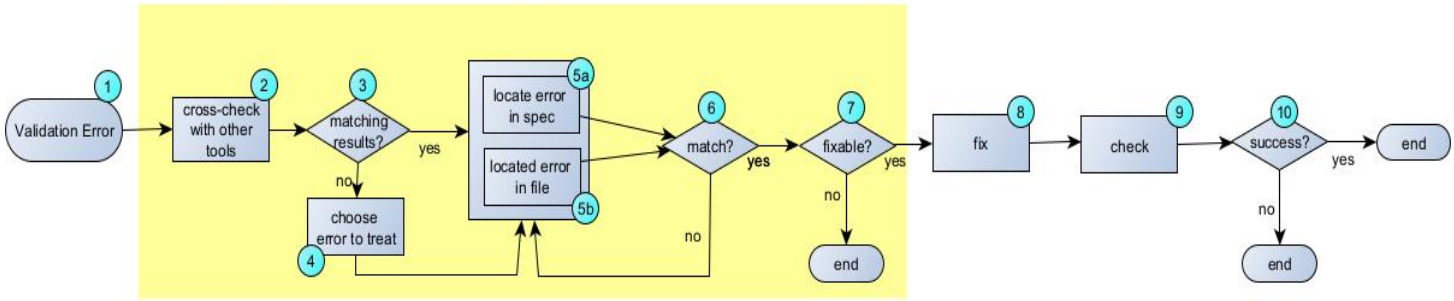


Figure 1: Basic overview of validation error treatment process. The yellow box includes the main analysis steps.

(2) Following a fixed path instead of ad-hoc processes will make it easier to identify gaps in the tools we use.

(3) An easy-to-compare documentation of what we do as a community is the prerequisite to questioning/checking/adapting our processes – it is the first step to next-level digital preservation.

But is a formalized description of what we do when validation fails even possible? This paper shall address exactly that question. After drafting a basic workflow of typical post-validation-error steps, this workflow is tested using two different file formats (TIFF and PDF) with two validation error examples each. In a second step, the basic workflow graphic is adapted according to the analysis outcome and the workflows usability is briefly discussed.

II. Related Work

Digital preservation practice rates "file format validation" as a key task of the ingest process. But what if file format validation fails? While the past decade has put forth new validators such as new JHOVE modules [5], veraPDF [6], DPFManager [7], MediaConch [8] or pdfcpu [9], little progress has been made in describing what to do when things go wrong. With few exceptions, that information largely stays among file format practitioners in our domain [10], [11]. Instead of promoting a broad discussion on these error messages within the community and aiming for joined solution approaches when it comes to handling invalid files, we often find ourselves questioning the process per-se [12].

While Gattuso and Goethals reported on a workflow used to assess and mitigate JHOVE validation errors at the National Library of New Zealand in 2017 [10], little work has been undertaken on formalizing a generic workflow for post-validation-error situations. Even the "Community Owned Workflows (COW)" section of the COPTR Wiki [13] includes only one validation-centric description, which does not really touch on error handling.

III. Methodology

For notation of the process, a simple flowchart style is used in order to make the diagrams easy to understand, thus allowing them to be of benefit to the widest audience possible. In a first step, a basic overview of the process is drafted. The single steps outlined are based on shared community experiences made in the past 10 years of digital preservation practice [1],[2], [10],[11],[15]. Figure 1 shows this basic overview.

The starting point of the workflow is a validation error message, while its ending point as indicated in Figure 1 is the result of the validation error treatment process. In the wider digital preservation process this might be a decision to accept or decline the file. However, capturing this decision is considered outside of the scope of this paper. The process is broken down into two larger categories – the analysis chain (see yellow box in Figure 1) and the "treatment" chain following the analysis. The steps are described in further detail in subsection III A "Definitions".

The basic overview shall be a starting point for documenting the post-validation process. In a next step, the workflow is tested against real-life use cases to see where it works and where it does not work. Of particular interest is the question of how the basic workflow works when it comes to different file formats and different "severity levels" of validation errors. Since file format characteristics differ widely, two different file format families are chosen as examples to check the workflow against. Both file formats are widely adopted, have an openly available

specification and more than one validation tool available. TIFF shall represent file formats that are of comparatively strict and simple structure; PDF shall represent file formats with a comparatively flexible and complex structure. For both file formats two different error messages are chosen – one "fixable" and one "not fixable" error each - to test the workflow description against. While within the scope of this paper all workflow descriptions start with a JHOVE error message, the workflow diagram is kept generic enough to work with any validation tool's error output.

A. Definitions

Before looking at the workflow diagram in further detail, a shared understanding of "validation" needs to be reached. The Community Owned digital Preservation Tool Registry (COPTR) classifies the function validation as a subset of the lifecycle stage ingest, describing it as "(...) the validation of digital files, typically against a file format specification" [32]. The dpc handbook has an even broader approach, stating that file format validation compares an instance of a file format to its expected behaviours [33]. More granular discussions of file format validation [2], [12], [15], [34] differentiate between different error levels of validation, such as "well-formed" and "valid" or "error" and "warning". Within the scope of this paper, (file format) validation is understood as any tool-based method to check a file format instance against a publically available description of the file format's syntax and semantics. This description can be in form of a full standard document, a format specification or a rule set, including a rule set of the validator itself.

The rest of this subsection gives a short overview of each of the workflow steps described in Figure 1 including their necessity and dependency. Necessity of a step depends on pathways chosen – e.g., step 4 ("choose error to treat") is optional, as it depends on more errors than one being present in the validation results.

Step 1: Validation Error (Mandatory)

Description: Starting point of the workflow; error can be from any tool used to validate the syntax and semantics of the file format

Prerequisite: Validation error message; access to the validation tool used; access to the file being validated

Step 2: Cross-check with other Tools (Optional)

Description: If other tools are available to check the validity of the file, these are run to cross-check and potentially gather further information; step is optional since further tools may not be available for all cases

Prerequisite: Availability of further tools to check validity of file

Step 3: Matching Results? (Optional)

Description: If different tool(s) are used to cross-check (step 2), tool outputs are compared to initial validation error message (step 1); the decision whether results match is not necessarily a straightforward task as terminologies may differ between tools

Prerequisite: Cross-check with other tools completed and results documented (step 2)

Step 4: Choose Error to Treat (Optional)

Description: If additional errors were found, a decision needs to be made which error is handled; in some cases errors may be connected to each other, leading to more than one error being handled in the following analysis and fix steps

Prerequisite: Additional tool(s) available (step 2) and additional validation errors found (step 3)

Step 5a: Locate Error in Spec (Mandatory)

Description: Validation tools check against a rule set which is derived from a standard or specification document for the file format; checking the validity of an error requires a comparison of the specification that is being checked against and the position in the file that triggered the error;

Prerequisite: Knowledge of and access to the documentation which the tool checks against (i.e., standards document, specification, schema)

Step 5b: Locate Error in File (Mandatory)

Description: The position in the file that triggered the error is typically referenced in the error message (e.g., via offset, tag name, chunk, etc); it can be accessed via tools like a hexeditor (for binary formats), an editor (for text based formats) or of a structure parsing tool such as itext RUPS [31] for PDF

Prerequisite: Information about section of file that triggered the error; access to an analysis tool like

hexeditor or editor; knowledge of how to navigate through the file formats structure

Step 6: Match? (Mandatory)

Description: Rationale for the error message are compared by checking the rule against the respective section of the file – this allows to check for false positives (validation tool errors); this step also forms the basis for understanding the impact of the tool, resulting in necessary information for a potential fix

Prerequisite: Rule that triggered the error and corresponding section in the file

Step 7: Fixable? (Mandatory)

Description: While some validation errors cannot be fixed, others can, but institutions may elect not to do so, e.g., because the error has no impact on rendering behavior; since the decision not to fix a file leads to the first end marker, step 7 is the last mandatory step in the workflow description

Prerequisite: Understanding of the error message and its impacts; tools / methods to conduct fix

Step 8: Fix (Optional)

Description: Repairing the file within the context of the validation error message (step 1); while some institutions may decide to discard the original after a successful fix, both versions (original and fixed) should be kept until the end of the workflow described here

Prerequisite: Knowledge of a method and availability of tools needed to fix the validation error within the file

Step 9: Check (Optional)

Description: Fixed files are cross-checked by rerunning the tools that produced the original validation error (step 1) as well as, if available, other validation tools (step 2); outcome of check determines whether workflow may need to start over again with a new validation error; in addition to validation checks, content-based integrity checks (where available) may be conducted to verify that actual content of the digital object was unchanged

Prerequisite: Original and repaired file for potential cross-checks; content-based integrity check tools or methods (where available)

Step 10: Success? (Optional)

Description: fixes can be successful or not – the two different outcomes typically serve as hooks for follow-up workflows within an archive (e.g., decline unfixable file)

Prerequisite: Understanding of impact of fix on digital object

IV. Analysis - TIFF

The following section describes processes for two different TIFF validation errors by using the basic flowchart description. The first error is one that can be fixed while the second error is one without a known remedy.

The starting seed validation error always stems from JHOVE v1.26 TIFF-hul 1.9.3[5]. Cross-checking is always completed with DPF Manager v3.5.1 [7] in full-check against Default mode as well as with ExifTool v12.44 [16]. The steps outlined in Figure 1 will be referenced by their respective numbers.

A. TIFF Use Case 1: TIFF-HUL-2 Tag 270 out of sequence

The error and handling described here is similar to that of a previously published blog-post [17]. The error has been reproduced in a file made available as *TIFF_Case-1.tif* in the dataset associated with the paper [18].

Step 1: Validation Error

The JHOVE validation error is *"TIFF-HUL-2: Tag 270 out of sequence."* As additional information, JHOVE gives the offset at which the error occurs: 178

Step 2: Cross-check with other tools

The error is cross-checked with DPF Manager and Exiftool. DPF Manager reports two errors and one warning: *IFD-0007 "Tags must be in strict ascending order" for IFD1* and *IDFE-0002 "Only 7-bit ASCII codes are accepted" for tag 270 ImageDescription"*. In addition, DPFManager lists one warning. However, DPFManager warnings are considered out of scope for this paper as the tool clearly differentiates between errors and warnings. The DPFManager output includes a reference to the part of the TIFF specification that is violated by the file – for both errors that is "*TIFF Baseline 6: Section 2: TIFF Structure. Page 15*". Exiftool (called with *-validate –warning –a* flags) returns two warnings: *"Entries in IFD0 are out of order"* and "*Tag ID 0x010e ImageDescription out of sequence in IFD0"*.

Step 3: Matching Results?

JHOVE'S *TIFF-HUL-2 "Tag 270 out of sequence"*, DPF Manager's *IFD-0007 "Tags must be in strict ascending order"* and Exiftool's *"Entries in IFD0 are out of order"* and *"Tag ID 0x010e ImageDescription out of sequence in IFD0"* appear to be matching results – although the different referencing of the IFD as IFD0 and IFD1 between DPF Manager and ExifTool are confusing. DPF Manager finds one additional error pertaining to a non-7 bit ASCII Code (in this case, a German Umlaut "ö") in tag 270.

Since the tool set we ran puts forth two different errors, we move along the "no" branch to Step 4.

Step 4: Choose Error to treat

We choose to neglect the DPF Manager Tag 270 non-7-bit-ASCII Character error for now and focus on the original JHOVE error, which was confirmed by DPF Manager and ExifTool.

Step 5A: Locate error in spec

Thanks to the detailed information returned by DPF Manager, we know exactly where to consult the TIFF specification. DPF Manager paraphrases the specification text for us, so we do not necessarily have to go look it up ourselves: *"The entries in an Image File Directory(IFD) must be in strictly ascending order by tag although the values which directory entry points need not be in any particular order"* [19],[20].

Step 5B: Locate error in file

JHOVE navigates us to two locations: while the offset is of little help here as the information in the binary cannot be understood easily, the tag number given in the error message itself is indeed helpful. With a tag viewer like ExifTool, we can extract the tags as they appear in sequence in the file and we can indeed see that tag number 270 is located between 305 and 317, so clearly not in ascending order.

Step 6: Match?

The error message "Tag 270 out of sequence" matches with what we have found in the file and can be verified. We therefore move along the "yes" branch to step 7.

Step 7 - 8: Fixable? & Fix

As described in [17], the error can be fixed with Exifool using the *–P –ImageDescription= -tagsfromfile @ -ImageDescription* flags. We move along the "yes" branch in step 7 and fix the file in step 8. This results in the creation of a new file with the correct tag order while maintaining all timestamp information. The fixed file is included as *TIFF_Case-1_fixed_1.tif* in the dataset associated with this paper [18].

Step 9 – End: Check & Success?

The success of the fix can be verified by re-running the file through JHOVE, DPF Manager and ExifTool as well as by manually re-inspecting the file as described in step 5B. In addition, the integrity of the image data can be verified by comparing the hash of the image data in the old file to that of the new file. This can be achieved with ImageMagick [30] using *identify –quiet –format "%#"*. We conclude the handling of this instance of TIFF-HUL 2 Tag 270 out of sequence by moving along the "Yes" branch to the workflow's "End" marker.

The case-specific workflow diagram is included as Appendix A1 to this paper.

While the workflow has been completed successfully for the specific JHOVE error used as a starting seed, we did encounter additional errors along the way. When checking the fixed file in Step 9, JHOVE returned the object as well-formed and valid, whereas DPF Manager continued to report the IDFE-0002 Error "Only 7-bit ASCII codes are accepted" for tag 270. While the error can be easily treated using the same workflow methodology, it imposes questions on how multiple error treatment should be reflected in the description. This question is elaborated on further in the Discussion section of this paper.

B. TIFF Case 2: TIFF-HUL-28 StripOffsets inconsistent with StripByteCounts

The error and handling described here is similar to that of a previously published blog-post [21]. The error has been reproduced in the file *TIFF_Case-2.tif* that is available in the dataset associated with this paper [18].

Step 1: Validation Error

The JHOVE validation error is "*TIFF-HUL-28: StripOffsets inconsistent with StripByteCounts: 1 != 55*". No further information is given.

Step 2: Cross-check with other tools

The error is cross-checked with DPF Manager and Exiftool. DPF Manager reports two errors: *IDFE-0002*

"Only 7-bit ASCII codes are accepted" for tag 270 ImageDescription as well as *STRIPS-0005 "Inconsistent strip lengths, the cardinality of stripoffsets and StripsBytesCount must match"*. In addition, DPFManager lists one warning, which has no impact on the file and shall be neglected in the scope of this paper. The DPF Manager output includes a reference to the part of the file format specification that is violated by the file – for *STRIPS-0005* that is *"TIFF Baseline 6: Section 8: Baseline Field Reference Guide, Page 40*". ExifTool (called with *-validate –warning –a* flags) returns one warning: *"Wrong number of values in IFD0 0x0111 StripOffsets".*

Step 3: Matching Results?

JHOVE's *TIFF-HUL 28 "StripOffsets inconsistent with StripByteCounts: 1 != 55"*, DPF Manager's *STRIPS-0005 "Inconsistent strip lengths, the cardinality of stripoffsets and StripBytesCount must match"* and ExifTool's "*Wrong number of values in IFD0 0x0111 StripOffsets"* appear to be matching results. DPF Manager finds one additional error pertaining to a non-7 bit ASCII-Code (in this case, a German "ß") in Tag 270.

Since the tool set we ran puts forth two different errors, we move along the "no" branch to Step 4.

Step 4: Choose Error to treat

We choose to neglect the Tag 270 non-7 bit ASCII character error and focus on the original JHOVE error.

Step 5A: Locate error in spec

Again, DPF Manager paraphrases the section of the specification: "*The cardinality of stripOffsets and the cardinality of StripsBytesCounts must be the same"*.

Step 5B: Locate error in file

We can locate the error by extracting the values from the binary data of the respective tags. This is possible by using ExifTool's *–b* option. Comparing the values in question, we see for *TIFF_Case-2.tif* that StripOffsets contains 1 value, StripByteCounts contains 55 values.

Step 6: Match?

The error message *"StripOffsets inconsistent with StripbyteCounts: 1 != 55"* can be verified in the file. We therefore move along the "yes" branch to step 7.

Step 7 - End: Fixable?

Unfortunately, the error means that the image data cannot be extracted from the file correctly [21]. Since the error is not fixable, we move along the "no" branch of step 7 and reach the end of the workflow here.

The case-specific workflow diagram is included as Appendix A2 to this paper.

While the workflow could be applied correctly to the use case, we discovered that we need to rely on outside information or knowledge when it comes to the question of fixable errors. This topic will be picked up later in the Discussion section of this paper.

V. Analysis – PDF

The following section describes processes for two different PDF validation errors against the basic flowchart description. The first error is one that can be easily fixed, while the second error is one without a known remedy.

The starting seed validation error always stems from JHOVE v1.26 PDF-hul 1.12.3 [5]. Cross-checking is always completed with pdfcpu (v.0.4.0. dev, *validation –mode strict*) [9] and qpdf (v. 9.1.1, *--check –verbose* options) [22]. The steps outlined in Figure 1 are referenced by their respective numbers.

A. PDF Use Case 1: PDF-HUL-137 No Pdf Header

The error of "junk data" before the header is common, especially in some older PDF files [23]. For this use case, an example of such a case discovered "in the wild" is used. The file is made available as *PDF_Case-1.pdf* via the dataset associated with this paper [18].

Step 1: Validation Error

The JHOVE validation error is *"PDF-HUL-137: No PDF Header"*. The Offset is given as 0.

Step 2: Cross-check with other tools

The error is cross-checked with pdfcpu and qpdf. Pdfcpu returns the validation error *"xRefTable failed: pdfcpu: headerVersion: corrupt pdf stream – no header version available"*, whereas qpdf returns no error.

Step 3: Matching results?

JHOVE's *PDF-HUL 137* matches the *"no header version available"* message thrown by pdfcpu. Since there is

only one error to treat, we move along the "yes" branch directly to step 5A.

Step 4: Choose error to treat

Step 4 is skipped as we moved directly to step 5A from the "yes" branch in step 3.

Step 5A: Locate error in spec

ISO 32000-1:2018 states in section 7.5.2 that *"The first line of a PDF file shall be a header consisting of the 5 characters %PDF- followed by a version number of the form 1.N where N is a digit between 0 and 7"* [24].

Step 5B: Locate error in file

Viewing the file in a hex editor, we can easily see that the first line is not the version info. There are 128 additional bytes before the %PDF-1.3 declaration.

Step 6: Match?

The error message *"No PDF Header"* can be verified via file inspection. We therefore move along the "yes" branch to step 7.

Step 7 – 8: Fixable? & Fix

The PDF can be fixed by removing the 128 bytes before the %PDF-1.3 declaration, as they are deemed "junk data" [25]. We move along the "yes" branch in step 7 and fix the error in step 8.

Step 9 – End: Check & Success?

The fix can be verified by rerunning the fixed file through JHOVE and pdfcpu. The file is now returned as well-formed and valid and we move along the "yes" branch in step 10 to the end of the workflow.

The fixed file is made available as *PDF_Case-1_fixed.pdf* in the dataset associated with this paper [18].

In addition, Adobe Acrobat Professional offers a compare tool, via which two PDFs can be compared and a difference report be generated. The case-specific workflow diagram is included as Appendix A3 to this paper.

B. PDF Use Case 2: PDF-HUL-38 Invalid Object Definition

The last use case is a difficult and unsolved PDF case. The issue has been previously discussed in a blog post [1]. It is chosen to test how well the proposed workflow diagram is applicable to complicated cases where it is hard to pinpoint the error. An example of such a case discovered "in the wild" is used and made available as *PDF_Case-2.pdf* via the dataset associated with this paper [18].

Step 1: Validation Error

The JHOVE validation error is *"PDF-HUL-38: Invalid Object Definition".* The Offset is given as 285259. For this particular test file, JHOVE throws another error as well: *PDF-HUL 87 "File header gives version as 1.3, but catalog dictionary gives version as 1.4".* Since the workflow description takes exactly 1 validation message as a starting point, and since PDF-HUL-87 is an error that can be neglected [26] we will focus on PDF-HUL-38.

Step 2: Cross-check with other tools

The error is cross-checked with pdfcpu and qpdf. Pdfcpu returns the validation error *"dereferenceObject: problem dereferencing object 91: pdfcpu: ParseObjectAttributes: can't find 'obj'"*. Qpdf shows a total of 28 error messages. 19 of those are *"object has offset 0"* for different objectIds, 5 of those comment on missing or incorrect entries in different objectIds, 2 deal with object 91 (*"expected n n obj"* and *"0 not found in file after regenerating cross-reference table"* and 2 of the error messages pertain to the document as a whole (*"file is damaged"* and *"attempting to reconstruct cross-reference table"*).

Step 3: Matching results?

It is hard to figure out whether the different error messages actually describe the same problem. Both, pdfcpu and qpdf have an error pointing to obj 91. Unfortunately, the JHOVE offset does not lead to obj 91 but to obj 194. None of the other tools have an error message pointing to obj 194. Qpdf is the only tool that reports the file as damaged. Since the error messages point to different sections of the file, we move along the "no" branch to step 4 to choose which error to treat and analyze in the following steps.

Step 4: Choose error to treat

Even though none of the other errors match with the JHOVE error message, we will stick with our starting seed message and attempt to treat the PDF-HUL-87 for obj 194 first.

Step 5A: Locate error in spec

According to the JHOVE error message documentation the error message occurs when a

keyword other than “obj” was found while paring an indirect object definition [27]. According to section 7.3.10 of the specification Object definitions need to follow the form “<obj.number> <obj.generation> obj” [24].

Step 5B: Locate error in file

We already navigated to the respective obj via the JHOVE offset in step 3 to cross-check the result against that of other validation tools. Indeed instead of the expected “194 0 obj” we find a “194 00obk”.

Step 6: Match?

The error message *“Invalid Object Definition”* can be verified via file inspection and we move along the “yes” branch to step 7.

Step 7-10: Fixable? – Success?

We can replace the faulty “194 00obk” in the HexEditor with “194 0 obj”. However, when checking the file with the same tools used in Step 2, qpdf and pdfcpu still show the same error messages as before whereas JHOVE now shows a different error message – *“PDF-HUL-66 Lexical Error”*. Since the file is mangled from a specific point onwards, the post-validation workflow process could be repeated countless times without reaching a successful result. The process is aborted here.

The case specific workflow diagram is included as Appendix A4 to this paper. The PDF with the applied fix for PDF-HUL-38 as per Step 8 is included in the dataset as *PDF_Case-2_fixed1.pdf* [18].

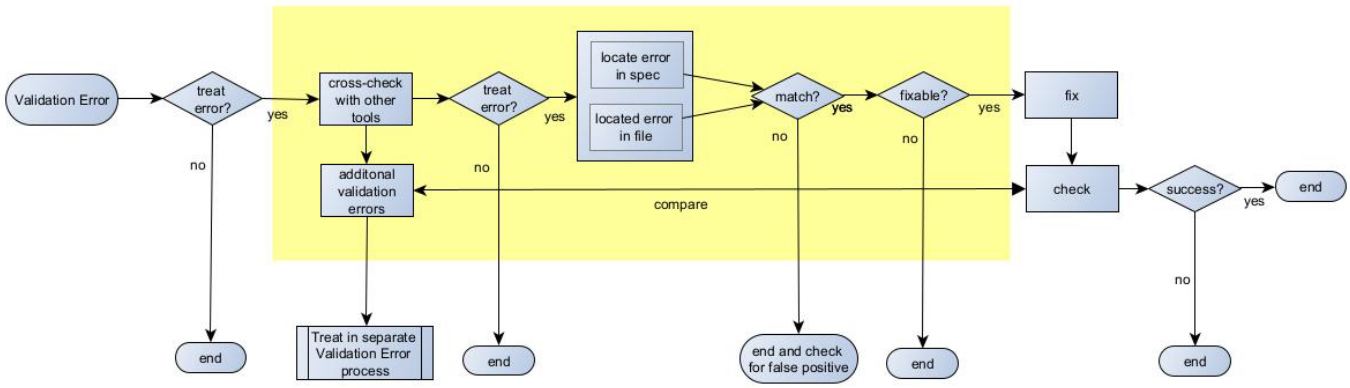
While the workflow described how the specific JHOVE error PDF-HUL-38 was treated, the digital object itself seems to have a bigger problem, which all reported errors are connected to. The question on whether this interdependency can be modeled in the workflow will be touched upon in the discussion section.

VI. Discussion

In testing the workflow diagram against the use cases we discovered several issues that should be reflected in an updated diagram version. This updated version is presented as Figure 2. A common strand was that for all use cases external information needed to be consulted to address the issues. This included error message description for JHOVE in the GitHub Wiki, blog posts or uncodified information known through practice. This is where tools can be easily improved – e.g., the JHOVE GUI could include pointers to the specification the same way that DPFManager does. This would especially help less experienced users in making the connection between the error and the expected file format syntax. As already discussed within the community, different error levels such as warnings in addition to errors would be a beneficial addition to JHOVE as well [28].

Both, the TIFF and PDF use cases included warnings or error messages which we did not treat (e.g., PDF-HUL-87 in the second PDF use case). While an in-depth discussion of these errors is out of scope for this paper, the question of how to model the decision not to treat an error is relevant to the diagram. But where in the workflow does that decision take place? While the first draft of the diagram presented in Figure 1 works on a “blank page” assumption, i.e., no knowledge of validators and error messages exist, the decision to ignore an error is always based on

Figure 2: Revised basic overview of validation error treatment process. The yellow box includes the main analysis steps.



existing knowledge. Often, errors such as the aforementioned PDF-HUL-87 or the "warning" messages included in the ExifTool output are specific to a tool. While a "warning" as opposed to an error could imply that no direct action is necessary, the actual decision not to treat the issue is always an individual one that may depend on an institutional policy. This binds the decision directly to a business rule, but also to the error message and its producing tool itself - the option not to treat the error therefore needs to be located at the beginning of the workflow. In some cases, however, the decision may depend on a "second opinion", i.e. a verification by a tool used to cross-check. Therefore, an optional additional "treat error" decision should be available after the cross-checking step. Figure 2 shows the updated diagram, with two "treat error" steps added directly after the validation error starting seed and again after the step "cross-check with other tools".

Knowing which errors not to treat goes hand in hand with knowing what errors should be treated. Once either an understanding of the error message itself and its correlation to the specification and file's actual syntax, or a solid trust in the validation tool has been established, the steps "locate error in spec", "locate error in file" and "match" may become unnecessary. These steps should therefore be optional. However, the same argument can be made for "cross-check with other tools", as this may become no longer needed once a high level of trust in one tool's ability is gained. Since the necessity of each step is therefore subjective, depending on the knowledge of those following the workflow, the mandatory / optional descriptors are removed from the updated workflow description.

But what if we come across multiple error messages within a file? And is the validation error message really the correct starting seed, or should the starting point instead be the digital object? In the first TIFF use case, JHOVE had only reported one error, whereas DPF Manager put forth a second error to be fixed. As shown in the use cases, validation error handling can become a complex task. In addition, we have to differentiate between error messages that are dependent on each other and those that are not. . The *"non-7-bit-ASCII"* error message introduced in TIFF Use case 1 is clearly an independent error message, whereas the 28 qpdf error messages found in the second PDF use case appear to depend on each other or on the same root cause. However, we want the workflow diagram to be an easy communication tool. Trying to model more than one validation error message at once in a diagram would make the diagram overly complex. Therefore, the decision is made to outsource additional error handling into new "Validation Error Handling" processes. The option to do so is added as a step resulting from the "cross-check with other tools". Since it might be helpful to understand if errors are dependent on each other or not, an optional "compare" connection was added to the diagram between the "additional validation errors" resulting from the initial cross-check and potential output from the "check" step post fixing.

Another issue exists with the conditional based on the match between the format specification and the error in file. In the first version of the diagram as presented in Figure 1, a successful matching leads to a fix, whereas an unsuccessful match leads to a re-evaluation of the specification and the error message. This could easily create endless loop if the connection simply does not exist, e.g. in case of a false positive returned by the validation tool. Instead of looping back to the comparison, a "no match" should exit the process and result in an evaluation of the validation error as a potential false positive.

Figure 2 presents the updated diagram with all changes included. The numbering and necessity of the steps has been removed due to aforementioned reasons.

VII. Conclusion and Outlook

The paper presented a first draft of a formalized methodology in form of a basic overview diagram for post-validation-error process steps. This first version was tested against four real-life use cases and updated based on the findings. As a general observation, the diagram outlines common steps but does not, of course, contain all the answers for what to do. However, having a structured documentation instead of having to sift through blog posts, wikis etc. to find the answer might make it easier for people to learn from and build on experiences made by others.

The introduction section listed three potential key benefits of such a formalized overview – one of those already proved achievable in form of recommendations for tool improvements made in the discussion section of this paper. Whether the diagram can be a vehicle for more effective

communication between practitioners and those just starting out and whether it can aid us in improving our processes remains to be seen.

A next step for this work is to collect community feedback and model more use cases on the updated version of the diagram. These use cases will then be included in the COPTR COW section. A long-term goal could also be to model validation error decisions in the Preservation Action Registry PAR [29].

1. REFERENCES

[1] Lindlar, M. "Trouble-shooting PDF validation errors – a case of PDF-HUL-38". Blogpost at the Open Preservation Foundation, Published on 27. November 2022. https://openpreservation.org/blogs/trouble-shooting-pdf-validation-errors-a-case-of-pdf-hul-38/

[2] Lehtonen, J. et al. "PDF Mayhem: Is Broken Really Broken?" In: Proceedings of the 15th International Conference on Digital Preservation, iPRES 2018, Boston, Massachusetts, USA, Sept. 24-28 2018. https://doi.org/10.17605/OSF.IO/FZXC9

[3] Digital Preservation Coalition. "iPRES 2022 – Bake Off: Full Menu – All You Can Eat". Recording of the Digital Preservation Bake Off Session held at the 18th International Conference on Digital Preservation, iPRES 2022, Glasgow, Scotland, Set. 12 – 16 2022. https://youtu.be/fj4Qg_Kj-xc

[4] OPF ICRFF Working Group. "International Comparison of Recommended File Formats" Version 1.2, April 2022. https://openpreservation.org/resources/member-groups/international-comparison-of-recommended-file-formats/

[5] JHOVE https://jhove.openpreservation.org/

[6] veraPDF https://verapdf.org/

[7] DPF Manager http://dpfmanager.org/

[8] Mediaconch https://mediaarea.net/MediaConch

[9] Pdfcpu https://pdfcpu.io/

[10] Gattuso, J., Goethals, A. "The Tip of the Validation Iceberg – Addressing JHOVE-based file validation warnings" In: Proceedings of the 14th International Conference on Digital Preservation, Kyoto, Japan, 25 – 29 September 2017. https://hdl.handle.net/11353/10.1424902

[11] Töwe, M., Geisser, F., Suri, R. "To Act or Not to Act – Handling File Format Identification Issues in Practice". In: Proceedings of the 13th International Conference on Digital Preservation, Bern, Switzerland, October 3 – 6, 2016. https://hdl.handle.net/11353/10.503183

[12] Whatley, P. "A valdediction for validation?" Blogpost at the DigitalPreservationCoalition. Published on 11 October 2018. https://www.dpconline.org/blog/a-valediction-for-validation

[13] COPTR Wiki: Community Owned Workflows (COW). https://coptr.digipres.org/index.php/Workflow:Community_Owned_Workflows

[14] Van der Knijff, J. "PDF processing and analysis with open –source tools". Blogpost at bitsgalore. Published on 06 September 2021. https://www.bitsgalore.org/2021/09/06/pdf-processing-and-analysis-with-open-source-tools

[15] Lindlar, M. Tunnat, Y. "How Valid is your Validation? A Closer Look behind the Curtain of JHOVE". In: International Journal of Digital Curation. Vol. 12, No. 2 (2017). https://doi.org/10.2218/ijdc.v12i2.578

[16] ExifTool. https://exiftool.org/

[17] Lindlar, M. "Troubles with TIFF: Tag 270 out of sequence". Blogpost at the Open Preservation Foundation. Published on 19 March 2020. https://openpreservation.org/blogs/troubles-with-tiff-tag-270-out-of-sequence/

[18] Lindlar, M. Dataset for this paper – currently at https://drive.google.com/drive/folders/1WmY0-nWvjq5aKwBdkvxu43_7h7iiLeU4?usp=sharing will be moved to Zenodo upon acceptance of paper

[19] DPS Manager. "Reference Documentation". http://dpfmanager.org/reference-documentation.html

[20] Aldus Developers Desk. "TIFF. Revision 6.0. Final" June 3, 1992.

[21] Lindlar, M. "Troubles with TIFF: StripOffsets inconsistent with StripByte Counts". Blogpost at the Open Preservation Foundation. Published on 12 April 2020. https://openpreservation.org/blogs/troubles-with-tiff-stripoffsets-inconsistent-with-stripbytecounts

[22] Qpdf https://qpdf.readthedocs.io/

[23] Lindlar, M., Tunnat, Y. "Time-travel with PRONOM: the 4th dimension of DROID". In: Proceedings of the 15th International Conference on Digital Preservation, iPRES 2018, Boston, Massachusetts, USA, Sept. 24-28 2018. https://doi.org/10.5281/zenodo.3517767

[24] ISO/TC 171/SC 2. "ISO 32000-1:2008 Document management – Portable document format – Part 1: PDF 1.7". 2008.

[25] OPF. "PDF-HUL-137". Error Message Wiki on github. https://github.com/openpreserve/jhove/wiki/PDF-hul-Messages-2#pdf-hul-137

[26] OPF. "PDF-HUL-87". Error Message Wiki on github. https://github.com/openpreserve/jhove/wiki/PDF-hul-Messages-2#pdf-hul-87

[27] OPF. "PDF-HUL-38". Error Message Wiki on github. https://github.com/openpreserve/jhove/wiki/PDF-hul-Messages#pdf-hul-38

[28] OPF. "Does JHOVE need a warning Message type?" JHOVE github issue #638 Discussion. https://github.com/openpreserve/jhove/issues/638

[29] Preservation Action Registries (PAR). https://parcore.org/

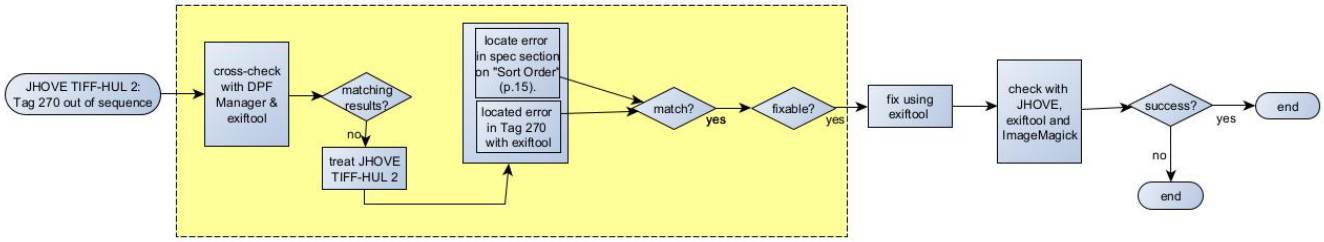
[30] ImageMagick. https://imagemagick.org/

[31] iText RUPS: https://github.com/itext/i7j-rups

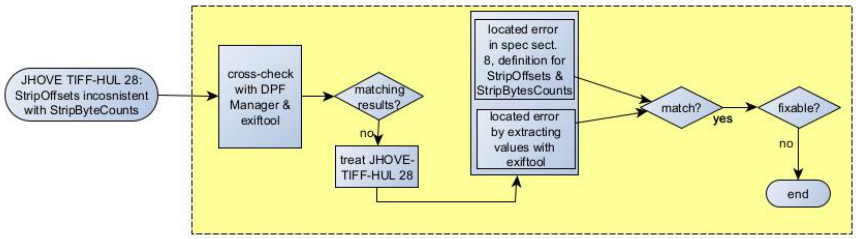
[32] COPTR function category"Validation": https://coptr.digipres.org/index.php/Validation

[33] dpc. "File formats and standards" In: Digital Preservation Handbook, 2nd Edition. 2015 https://www.dpconline.org/handbook .

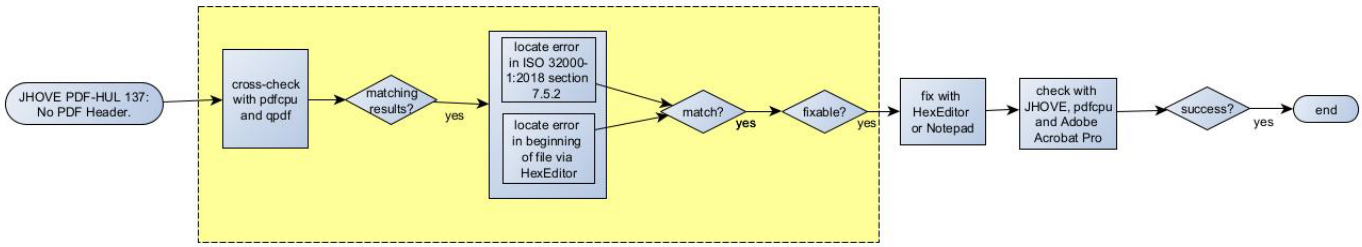
Appendix A1: Workflow for TIFF Use Case 1 – TIFF-HUL-2 Tag out of Sequence



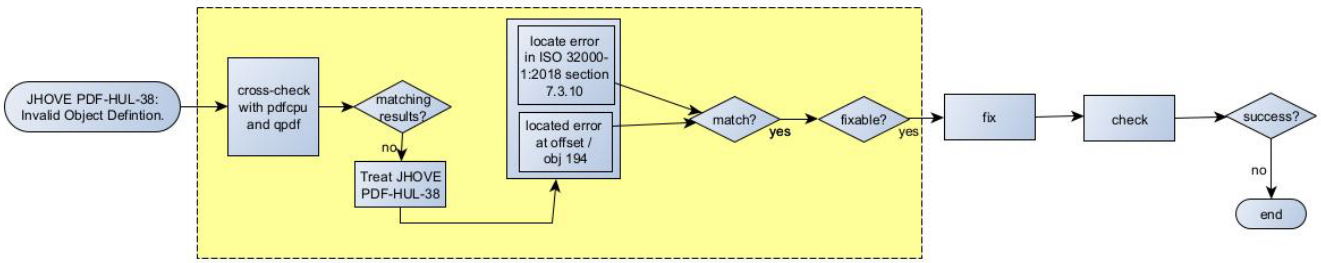
Appendix A2: Workflow for TIFF Use Case 2 – TIFF-HUL-28: StripOffsets inconsistent with StripByteCounts



Appendix A3: Workflow for PDF Use Case 1 – PDF-HUL 137: No PDF Header



Appendix A4: Workflow for PDF Use Case 2 – PDF-HUL-38: Invalid Object Definition



MULTILINGUAL LABELS FOR DIGITAL PRESERVATION

Katherine Thronton
Yale University Library
United States
katherine.thornton@yale.edu
0000-0002-4499-0451

Kenneth Seals-Nutt
Yale University Library
United States
kenneth.seals-nutt@yale.edu
0000-0002-5926-9245

Abstract – We introduce a technique for finding multilingual translations for lists of words using technologies of the Semantic Web. We present four subsets of data from Wikidata and Wikipedia as sources of multilingual labels. Our sample dataset consists of seven terms related to digital preservation. We compare the number of labels we can source for these terms from other human languages via SPARQL queries using the Wikidata Query Service. After discussing the composition of each subset, we detail their advantages and disadvantages. Providing multilingual la- bels as additional access points for resources such as on- tologies, vocabularies and user interfaces for applications increases the relevance of these resources to a larger per- centage of the global population. Increasing multilingual access promotes inclusion for a broader range of people, which leads to greater diversity in the digital preservation community.

Keywords – Wikidata, Semantic Web, multilingual data, knowledge graph subsets

Themes – Digital Accessibility, Inclusion, and Diversity, We're All in this Together

I. INTRODUCTION

Digital preservation is an international field made up of practitioners from all parts of the globe. Resources such as ontologies, vocabularies, and applications relevant to the work of digital preservation are frequently monolingual. English is used as the primary language for many resources. Such monolingual resources restrict their audience to people who have knowledge of English, while all others are excluded. Increasing the number of multilingual access points within such resources promotes equity and broadens the diversity of audiences who can benefit from the field of digital preservation. Maintainers of monolingual resources are faced with budgetary constraints, and may argue that expanding multilingual access is too expensive to be practical. Translations created by human experts are expensive, perhaps we can leverage the technologies of the Semantic Web to source multilingual labels as a cost-effective alternative.

We describe four subsets of Wikidata that people may find useful for sourcing multilingual labels. We created a sample data set to test for multilingual label coverage, and we describe the results of consulting four subsets of Wikidata for each term in the sample data set. After describing each subset, we discuss advantages and disadvantages of each. We introduce an interactive application that we created to browse each set of multilingual labels. We offer this work as a demonstration of how communities of editors who contribute to the projects of the Wikimedia Foundation have created a valuable multilingual knowledge graph. The fact that all of the data in Wikidata is free for anyone to reuse for any purpose makes this a shared international resource. Times of international crises such as global pandemic, or armed conflict, reinforce the importance of striving to make resources more equitably available. Providing cost-effective strategies for sourcing multilingual labels is a pathway to promoting equity through increasing multilingual access.

II. WIKIDATA

Wikidata is a project of Wikimedia Deutschland, the German chapter of the Wikimedia Foundation. The Wiki- data community launched this public knowledge base of structured data in 2012. The architecture of Wiki- data was designed from the outset to support multilingual content [1]. The

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

Wikidata knowledge base contains Items that can be connected to literal values, or to other Items, through the use of Properties [2]. The Wikidata community has added more labels in English than any other supported language, but there are dozens of additional languages for which the Wikidata community has also added many labels [3]. The work of the members of the Wikidata community to add multilingual labels to Items and Properties has resulted in a corpus of equivalent labels across hundreds of human languages.

III. Related Work

Wikidata is an example of a collaboratively-created knowledge graph [4]. After ten years of existence, Wiki- data is well-recognized as valuable source for reusable data [5]. Researchers have leveraged multilingual content from Wikidata for various applications. For example, multilingual content from Wikidata has been used to power a question-answering platform [6], and has been used to generate article placeholders for encyclopedias [7]. The challenge of organizing access terms for multi- lingual digital content has been addressed by language- independent mappings drawn from the Semantic Web [8]. Multilingual access is necessary for national con- texts in which multiple languages are supported [9]. Due to the fact that the digital preservation community is an international community, it is clear that we need to pro- vide access to our applications and resources in a wide range of human languages [10].

We sampled several subsets of Wikidata for this work. Wikidata subsets are portions of the Wikidata knowledge graph [11]. The size of Wikidata makes it desirable to reuse a subset, as it is time-consuming to process and host the entire Wikidata graph. Often subsets are focused around a particular type of data, or a specific domain. We identified subsets of Wikidata related to seven sample terms, and wrote SPARQL queries to ex- tract associated data from Wikidata. Subsets of Wikidata may be extracted by a variety of software tools, or via the Wikidata Query Service. An overview of tools available to extract subsets from Wikidata is provided by [12]. Researchers have also explored memory-efficient techniques that allow for larger subsets to be extracted more quickly than techniques that use Wikidata's SPARQL end- point [13].

Researchers and practitioners approach the translation of ontologies, vocabularies, and other term-based resources using a variety of methods. One approach is to extend an ontology with multilingual labels [14]. A successful tri-lingual project is described in [15]. Others have explored using Wikipedias to generate translations [16]. Our approach differs in that we combine translations from Wikipedias along with additional multilingual content from Wikidata, thus extending coverage from additional human languages.

IV. Sample Data Set

We selected seven terms related to digital preservation to create a sample data set[1]. The terms we included are: file format, checksum, operating system, data integrity, software, license, and reproducibility. We chose these terms because of their relevance to digital preservation work activities. We then searched the Wikidata knowledge base to gather the Qids for the relevant Wiki- data items. Using the Qids for these terms, we wrote SPARQL queries to identify multilingual labels for these terms. The Wikidata items served as the basis for three of the subsets: the Wikidata Item Labels, the Wikipedia Article Titles, and the Wikidata Lexemes. To find our fourth subset, we searched the Property namespace for our terms to retrieve the Property Labels.

We created an interactive application that presents the labels available in each of the four subsets for each of the words in our sample dataset[1]. This application allows anyone to quickly compare the language cover- age per subset for each term. For example, in Table 1, we see a visualization of the languages (represented by their ISO codes[2]) color-coded if we have a label in that language, and without color if we do not. As the user hovers over a language, the label itself will be dis- played alongside the name of the language. There are drop-down menus that users can select from in order to switch between terms from the sample data set and to switch between the four subsets. The layout of languages is consistent across the different views, allowing visual comparison of the overlap between

[1] The webapp that includes the interactive table is available at https://wikidp-research.k2.services/multi-lingual-table.

[2] https://www.iso.org/iso-639-language-codes.html

subsets. In Figure 2, we see the labels available from the Wikidata item for 'software' (Q7397).

Figure 1: Labels for 'software' from Wikipedia Article titles, as seen in https://wikidp- research.k2.services/multi-lingual-table

Figure 2: Labels for 'software' from Wikidata, as seen in https://wikidp-research.k2.services/multi-lingual-table

V. Reusing Multilingual Content from Wikimedia Projects

The human editors working to create and extend content in Wikimedia projects are constantly working to improve the quality of information across the projects. The large number of people who view and edit this con- tent help to remove errors and ensure accuracy. Content in Wikidata is published under the Creative Com- mons Zero license, meaning that data in Wikidata is free for anyone to reuse for any purpose. The Wikidata SPARQL endpoint[3] is a public endpoint that anyone can use to request data from Wikidata [17]. No credentials are needed to run queries on Wikidata's SPARQL end- point, making this a convenient method of data retrieval. We introduce four subsets in this section: Wikipedia Article Titles, Wikidata Item Labels, Wikidata Property Labels and Wikidata Lexemes. Data from each of these subsets is available from the Wikidata Query Service.

A. Article Names per Language Version of Wikipedia

One early layer of data in Wikidata is that of interwiki links. Interwiki links connect articles that describe a topic among the different language versions of Wikipedia. These interwiki links are now stored in Wikidata, meaning that Wikidata items are connected to corresponding Wikipedia articles [1]. The titles of the articles in the different language versions are a potential source of multilingual labels for these terms. New language versions of Wikipedia are still being created. There are more than three hundred versions of Wikipedia [18]. Hypothetically, if every language version were to have an article about file formats, we would then have hundreds of multilingual labels from the set of article titles. We wrote SPARQL queries to return the article titles from each of these language versions of articles about file formats.

For example, there are 44 versions of Wikipedia that have an article about file formats, as seen in Figure 3. We can retrieve all of these article titles and consider them multilingual label candidates. The largest number of la- bels from the Wikipedia Articles subset is available for 'operating system' with 150 potential labels. This is due to the fact that more Wikipedia communities have writ- ten articles about 'operating system' than about any of the other terms from our sample set. Only twenty-seven language versions of Wikipedia have articles about 'reproducibility'. This is likely due to the frequency of usage of these terms, and thus relevance for an average contributor to Wikipedia. The Google Books Ngram Viewer[4], which presents occurrence data for search terms as seen in the corpus of Google Books, demonstrates that 'operating system' is found more frequently than 'reproducibility' between the years 1960-2019, as seen in Figure 4.

[3] https://query.wikidata.org/

[4] https://books.google.com/ngrams/

Figure 3: Interwiki links for the different language ver- sions of Wikipedia that contain an article titled 'file for- mat'.

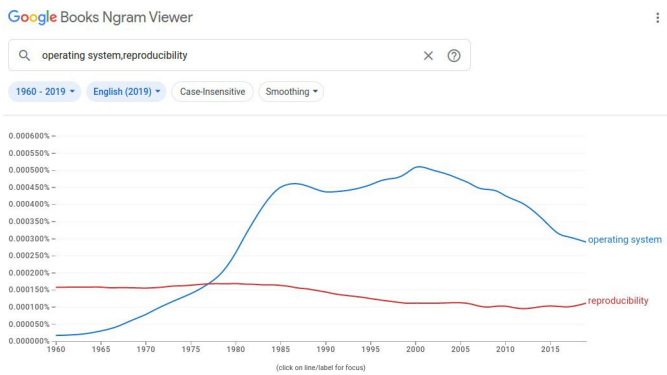


Figure 4: Google Books Ngram Viewer results for 'oper- ating system' and 'reproducibility' from 1960-2019

In Table 1 we see the count of labels available in the Wikipedia Article Title subgraph for each of the terms in our sample data set. An advantage of sourcing labels from this subset is that article titles have high visibility within Wikipedias, thus these labels are likely to be corrected very quickly if they are vandalized or require improvement. A disadvantage of sourcing labels from this subset is that new articles are created on a relatively slow timeline, thus this subset is likely to grow slowly. If we compare the work involved in writing a new article in Wikipedia with adding a label to a Wikidata item, writing a new article requires substantially more effort.

Term	Count Wikipedia Article Titles
file format	44
checksum	37
operating system	150
data integrity	30
software	131
license	65
reproducibility	27

Table 1: Count of Labels from Wikipedia Article Titles

B. Multilingual Item Labels from Wikidata

The designers of the data model for Wikidata intended it to be a multilingual knowledge base [1]. Each item in Wikidata has a Qid identifier composed of the letter Q and numbers. These Qids are designed to avoid privileging one human language over others supported by the knowledge base. The Wikidata data model sup- ports labels in more than three hundred languages [3]. Wikidata editors add labels in many different languages. Over time, the set of all labels for a particular item becomes a very useful set of translations.

An advantage of sourcing labels from this subset is that Wikidata labels are added at a faster pace than new articles are created, this this subset is likely to grow more quickly over time. In order to add a label, users type the string into the user interface in the area designated for the language of choice, and then press 'publish' to contribute the content. Wikidata item labels are seen by many editors, as well as by many people who reuse data from Wikidata, thus these labels are likely to be updated quickly if they require improvement. Multilingual labels are an aspect of Wikidata that some editors monitor closely [3].

In Table 2 we see the count of labels available in the Wikidata Items subgraph. To date the terms from our sample data set with the largest number of available la- bels are ‘software’ and ‘operating system’, each with la- bels in more than one hundred human languages. The terms ‘checksum’ and ‘reproducibility’ have fewer available labels.

C. *Multilingual Property Labels from Wikidata*

Four of the terms in our sample dataset are related to properties in Wikidata. Wikidata properties are predicates that describe how items are related to one another. Properties are also modeled to accommodate la- bels in all languages supported by Wikidata. Some members of the Wikidata community specialize in working on property labels [19]. We can consult the subgraph of property labels for our sample dataset to see if there are any additional labels in languages not yet covered by the other subsets. While this may result in some redundant labels, as we would expect the labels to be the same for the item and the property, there could be some additional languages that have coverage in the property label subgraph. For example, in Figure 5, we see some of the labels available for the Wikidata item ‘checksum’ include labels from Thai and Ukrainian, but not Turkish. In contrast, in Figure 6, we see that a Turkish label is available.

Term	Count Wikidata Item Labels
file format	55
checksum	39
operating system	105
data integrity	33
software	103
license	70
reproducibility	35

Table 2: Count of Labels from Wikidata Item Subgraph

A disadvantage of sourcing labels from this sub-graph is that there are a limited number of properties in Wikidata. There are currently more than 10,000 proper- ties in Wikidata, but more than 100,000,000 items. Thus there are many terms that will not be found among properties. In Table 3 we see the count of labels available in the Wikidata Property label subgraph. For terms in our sample data set that are not related to a Wikidata property, we recorded N/A in the table.

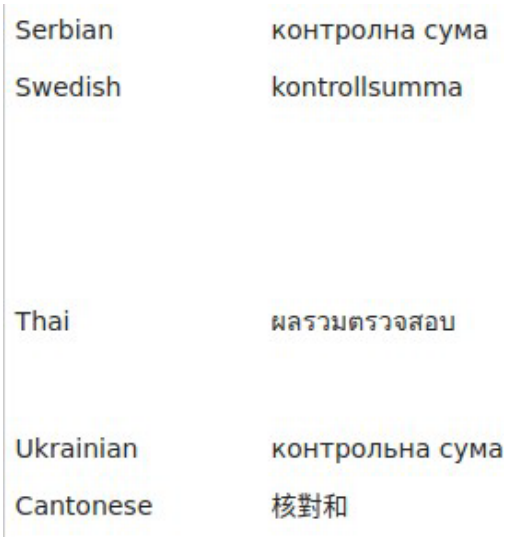


Figure 5: Some of the labels available for the Wikidata item ‘checksum’ (Q218341)

Figure 6: Some of the labels available for the Wikidata property ‘checksum’ (P4092)

Term	Count Wikidata Property Labels
file format	38
checksum	26
operating system	69
data integrity	N/A
software	N/A
license	76
reproducibility	N/A

Table 3: Count of Labels from Wikidata Property Sub- graph

D. *Multilingual Property Labels from Wikidata*

Wikidata also contains detailed linguistic data in the Lexeme namespace. Community members create lexemes, forms, and senses in the Lexeme

namespace following the data model for lexicographical data [20]. The Lexeme namespace, namespace L, was created in 2018 [21]. Wikidata has a property that is used to connect Lexeme senses to corresponding Wikidata items. The property has the English label 'item for this sense' and is P5137. Through the use of this property, the Wiki- data items from our sample data set can be connected to lexeme senses. In Figure 7, we see the lexeme 'soft- ware' (L1135). In the section of the page with the heading 'Senses' we see that the property 'item for this sense' has the value 'software' which is the Wikidata item identified with Q7397.

The graph of lexeme senses and their connections to Wikidata items is likely to increase in size over time. Currently, for this sample data set there are zero lexeme senses for 'checksum' and 'data integrity'. This is likely due to the fact that these concepts are domain-specific, and relatively infrequently used by people who are not engaged with the domain of computing.

A useful tool for searching for lexemes is Ordia [22]. Ordia can be used to search for lexemes and provides overviews of connections between lexemes and other content. Wikidata editors to the Lexeme namespace have already contributed more than half a million lexical entries [23].

An advantage of sourcing labels from this subset is that it is likely to grow in the future. As more editors use the property 'item for this sense' P5137 to connect Lexeme senses to Wikidata items, this subgraph will grow. Lexemes are connected to external identifiers related to etymology, dictionaries and other linguistic resources. Depending on the type of resource for which you are sourcing multilingual labels, pointers to additional linguistic information may also be helpful. Another advantage is that the data model for lexicographic data in the L namespace accommodates the use of references. Lexemes can also be connected to authoritative sources from which information was sourced. For example, in Figure 9, we see that the Swedish noun 'licens' is sourced back to Svenska Akademiens Ordbok using the property 'described by source' in Wikidata's Lexeme namespace. This increases the value of labels sourced from the lexeme subset as they may also include provenance information.

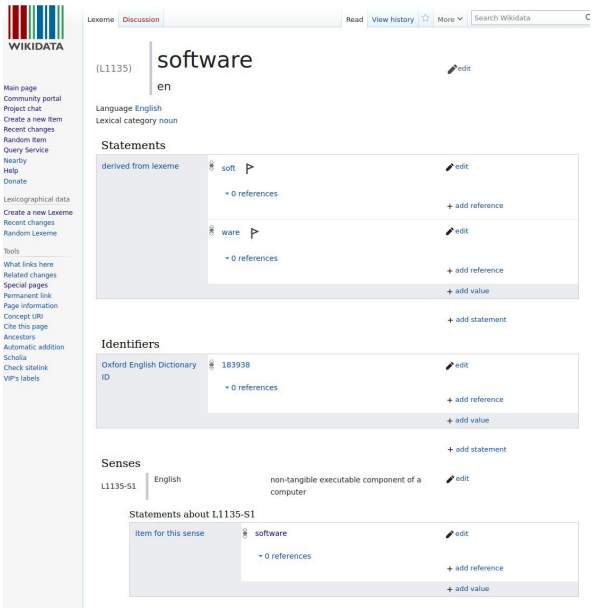


Figure 7: Lexeme L1135 'software' in Wikidata

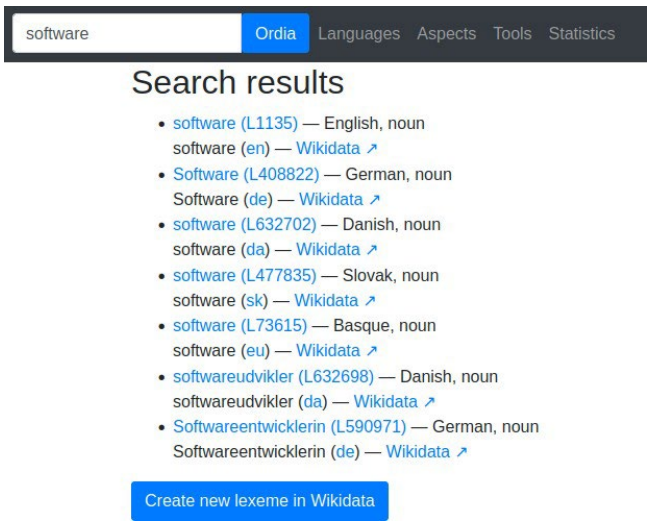


Figure 8: Screenshot from the Ordia application showing a search for 'software' in Wikidata's L namespace.

Figure 9: The Swedish noun 'licens' is connected to Sven- ska Akademiens Ordbok using the property 'described by source' in Wikidata's Lexeme namespace.

The disadvantage of sourcing labels from this subset is that there are not as many editors contributing edits to the Lexeme namespace in Wikidata as there are editors who contribute to other namespaces.

In Table 4 we see the count of labels available in the Wikidata Lexemes subgraph. The terms 'software' and 'license' currently have the largest number of lexeme senses that have been connected to their Wikidata items. As encouragement for more Wikidata editors to familiarize themselves with the Lexeme namespace, people organize weekly challenges with a topical focus. For example, one recent lexeme challenge was focused on software[5] and another on computing[6]. These challenges are announced via Wikidata-related communication channels. We anticipate that as more editors learn about the lexeme namespace this subgraph is likely to increase in size.

Term	Count Wikidata Lexeme Senses
file format	1
checksum	0
operating system	3
data integrity	0
software	16
license	23
reproducibility	1

Table 4: Count of Labels from Wikidata Lexeme Sub- graph

VI. Discussion

The multilingual labels available via the Wikidata Query Service could be of value to people who are looking to source translations for terms in an ontology, vocabulary, glossary or for text in the user-interface of an application. While the number of available labels varies across terms, the open licensing of the data and the accessibility of the data via the Wikidata Query Service make this an attractive cost-free alternative to hiring translators for groups with limited budgets.

Looking at the count of labels available for the terms in our sample data set it is clear that editors have added more labels for 'software' and 'operating system' than the other terms. This is likely due to the high levels of awareness many editors have for these terms. The other terms in the sample data set are more specialized, and thus may be less familiar to editors. To date, editors have added the fewest number of labels for the terms 'reproducibility' and 'data integrity'. Fewer editors may be familiar with these terms, or have use cases that would lead them to edit these items.

The webapp[7] we created to complement this paper allows viewers to see each label in the context of the set of supported languages. Not only can you get a sense of how many labels are available per ter for each subset, it is also possible to see each label if you hover over the colored language blocks in the webapp.

Members of the Wikidata community are motivated to contribute for many different reasons. There is no group or individual dictating how others should con- tribute to the project [24]. Different subsets of Wiki- data have different numbers of labels for the terms in our sample data set because there is no coordination of how work is accomplished, other than ad hoc decisions among editors to collaborate. This is consistent with the theoretical work describing peer-productions systems [25], [26].

As more people with digital preservation expertise decide to become editors of projects of the Wikimedia Foundation, it is possible that editors from our international community of practice could contribute more la- bels in additional languages to items, properties, and lexeme senses to Wikidata or contribute new articles in additional language versions of Wikipedia. Such contributions would benefit anyone interested in reusing data from Wikidata or Wikipedia. Guidance related to con-tributing to Wikidata tailored to the digital preservation community is described in [27]. Leveraging the infrastructure of the projects of the Wikimedia Foundation for collaboration is a strategy for that supports users from many different language contexts to benefit [28].

VII. Conclusion

People who create or maintain vocabularies, ontologies, applications or other projects may require multilingual labels for concepts in their systems. The cost of paying for translations into multiple languages can quickly add up, and may be beyond the budgetary constraints of many projects. Not only is there a multilingual knowledge graph that is free to reuse, exploring the multilingual data in the

[5] https://dicare.toolforge.org/lexemes/challenge.php?id=52

[6] https://dicare.toolforge.org/lexemes/challenge.php?id=28

[7] The webapp that includes the interactive table is available at https://wikidp-research.k2.services/multi-lingual-table.

projects of the Wikimedia Foundation is an approachable task using the Wikidata Query Service. The Wikidata Query Service provides multiple options for downloading results in formats such as CSV, JSON, or HTML, they also provide code snipits for reusing queries within external applications, as seen in Figure 10. Once a subset has been identified, either through SPARQL queries or a ShEx schema in the Entity Schema namespace, results may be consulted again at a later point to determine if additional data is added by the Wikidata community over time. Subsets can be reused in other applications, to enrich ontologies, vocabularies, or within other resources where multilingual labels are needed.

URL HTML Wikilink PHP JavaScript (jQuery)
JavaScript (modern) Java Perl Python Python (Pywikibot)
Ruby R Matlab listeria mapframe

```
# pip install sparqlwrapper
# https://rdflib.github.io/sparqlwrapper/

import sys
from SPARQLWrapper import SPARQLWrapper, JSON

endpoint_url = "https://query.wikidata.org/sparql"

query = """SELECT ?label ?languageCode WHERE {
  hint:Query hint:optimizer "None".
  ?article schema:about wd:Q7397;
    schema:name ?label;
    (schema:isPartOf/wikibase:wikiGroup) "wikipedia".
  hint:Prior hint:gearing "forward".
  ?article schema:inLanguage ?languageCode.
}"""


def get_results(endpoint_url, query):
    user_agent = "WDQS-example Python/%s.%s" %
(sys.version_info[0], sys.version_info[1])
    # TODO adjust user agent; see https://w.wiki/CX6
    sparql = SPARQLWrapper(endpoint_url, agent=user_agent)
    sparql.setQuery(query)
    sparql.setReturnFormat(JSON)
    return sparql.query().convert()


results = get_results(endpoint_url, query)

for result in results["results"]["bindings"]:
    print(result)
```

Figure 10: Python code snipit available from the Wiki- data Query Service

Sourcing labels from multiple subsets of Wikidata increases the breadth of languages that can be covered. People who are committed to holding themselves ac- countable to the values of accessibility, inclusion, and diversity may want to consider sourcing multilingual labels for resources in the domain of digital preservation from the projects of the Wikimedia Foundation. Some members of the digital preservation community may wish to contribute labels in their own languages for these terms, or for any other items or properties in Wikidata related to digital preservation, in order to improve and extend the knowledge graph.

We offer the techniques described in this paper for identifying potential subsets of multilingual data as strategies that others in the digital preservation community may find helpful. Investigating the multilingual label inventory from projects of the Wikimedia Foundation via the Wikidata Query Service could reduce or eliminate the need to source multilingual translations from other, more expensive, sources. As we are all in this together, let's support one another in our shared goals of increasing multilingual access points in projects and tools used by the digital preservation community.

VIII. Acknowledgements

We thank the communities of contributors to the projects of the Wikimedia Foundation. Contributions to projects of the Wikimedia Foundation made this re- search possible.

References

[1] D. Vrande?i?, "Wikidata: A new platform for collab- orative data collection," in Proceedings of the 21st International Conference Companion on World Wide Web, ACM, 2012, pp. 1063-1064.

[2] W. Community, Wikidata:data model, 2023. [On- line]. Available: https : / / www . wikidata . org / wiki/Wikidata:Data_model.

[3] L.-A. Kaffee, A. Piscopo, P. Vougiouklis, E. Sim- perl, L. Carr, and L. Pintscher, "A Glimpse into Babel: An Analysis of Multilinguality in Wikidata," in Proceedings of the 13th International Symposium on Open Collaboration, ser. OpenSym '17, Galway, Ireland: ACM, 2017, 14:1-14:5, isbn: 978-1-4503- 5187-4. doi: 10.1145/3125433.3125465. [Online]. Available: https://doi.org/10.1145/3125433. 3125465.

[4] A. Hogan, E. Blomqvist, M. Cochez, et al., "Knowl- edge graphs," Synthesis Lectures on Data, Seman- tics, and Knowledge, vol. 12, no. 2, pp. 1-257, 2021.

[5] L. Jarnac and P. Monnin, "Wikidata to bootstrap an enterprise knowledge graph: How to stay on topic?" In Proceedings of the 3rd Wikidata Workshop 2022. [Online]. Available: https://ceur-ws.org/ Vol-3262/paper16.pdf.

[6] T. Pellissier Tanon, M. D. de Assunção, E. Caron, and F. M. Suchanek, "Demoing platypus-a multilin- gual question answering platform for wikidata," in The Semantic Web: ESWC 2018 Satellite Events: ESWC 2018 Satellite Events, Heraklion, Crete, Greece, June 3- 7, 2018, Revised Selected Papers 15, Springer, 2018, pp. 111-116.

[7] L.-A. Kaffee, H. ElSahar, P. Vougiouklis, et al., "Mind the (language) gap: Generation of multilingual wikipedia summaries from wikidata for article- placeholders," in The Semantic Web: 15th Inter- national Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings 15, Springer, 2018, pp. 319-334.

[8] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar, and J. McCrae, "Chal- lenges for the multilingual web of data," Journal of Web Semantics, vol. 11, pp. 63-71, 2012.
[9] M. Bremer-Laamanen and J. Stenvall, "Selection for digital preservation: Dilemmas and issues," in Managing preservation for libraries and archives, Routledge, 2018, pp. 53-65.
[10] T. Evens and L. Hauttekeete, "Challenges of digi- tal preservation for cultural heritage institutions," Journal of Librarianship and Information Science, vol. 43, no. 3, pp. 157-165, 2011.
[11] J. E. L. Gayo, Creating knowledge graphs subsets us- ing shape expressions, 2021. arXiv: 2110 . 11709 [cs.DB].
[12] J. E. Labra-Gayo, A. C. G. Cavazos, A. Waagmeester, et al., "Enhancement and reusage of biomedical knowledge graph subsets," 2022.
[13] P. Nguyen and H. Takeda, "Wikidata-lite for knowl- edge extraction and exploration," arXiv preprint arXiv:2211.05416, 2022. [Online]. Available: https://arxiv.org/pdf/2211.05416.
[14] M. Espinoza, A. Gómez-Pérez, and E. Mena, "En- riching an ontology with multilingual information," in The Semantic Web: Research and Applications: 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008 Pro- ceedings 5, Springer, 2008, pp. 333-347.
[15] S. Niininen, S. Nykyri, and O. Suominen, "The fu- ture of metadata: Open, linked, and multilingual- the yso case," Journal of Documentation, 2017.
[16] A. Conde, A. Arruarte, M. Larrañaga, and J. A. Elor- riaga, "How can wikipedia be used to support the process of automatically building multilingual do- main modules? a case study.," Information Process- ing & Management, vol. 57, no. 4, p. 102 232, 2020.
[17] S. Malyshev, M. Krötzsch, L. González, J. Gonsior, and A. Bielefeldt, "Getting the most out of wikidata: Semantic technology usage in wikipedia's knowl- edge graph," in International Semantic Web Confer- ence, Springer, 2018, pp. 376-394.
[18] Meta, List of wikipedias meta, discussion about wikimedia projects, [Online; accessed 8-October- 2022], 2022. [Online]. Available: {https://meta.wikimedia.org/w/index.php?title=List_of_Wikipedias&oldid=23800107}.
[19] T. Pellissier Tanon and L.-A. Kaffee, "Property label stability in wikidata: Evolution and convergence of schemas in collaborative knowledge bases," in Companion Proceedings of the The Web Conference 2018, 2018, pp. 1801-1803.
[20] W. community, Wikidata:lexicographical data/documentation, Online; accessed 9- March-2023, 2023. [Online]. Available: https :/ / www . wikidata . org / wiki / Wikidata : Lexicographical_data/Documentation.
[21] F. Nielsen, "Lexemes in wikidata: 2020 status," in Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020), 2020, pp. 82-86.
[22] F. Å. Nielsen, "Ordia: A web application for wiki- data lexemes," in European Semantic Web Confer- ence, Springer, 2019, pp. 141-146.
[23] Ordia, Ordia statistics, [Online; accessed 13- October-2022], 2022. [Online]. Available: {https://ordia.toolforge.org/statistics/}.
[24] C. Müller-Birn, B. Karran, J. Lehmann, and M. Luczak-Rösch, "Peer-production system or collab- orative ontology engineering effort: What is wiki- data?" In Proceedings of the 11th International Symposium on Open Collaboration, ACM, 2015, p. 20.
[25] Y. Benkler, "Coase's penguin, or, linux and the nature of the ?rm," Yale Law Journal, pp. 369-446, 2002.
[26] Y. Benkler, A. Shaw, and B. M. Hill, "Peer production: A modality of collective intelligence," Collective Intelligence, 2013.
[27] K. Thornton, "Wikidata for Digital Preservation- ists," en, Dec. 2, 2021, 9 pp. doi: 10.7207/TWGN21-19. [Online]. Available: http://dx.doi.org/10. 7207/twgn21-19.
[28] J. Samuel, "Towards understanding and improving multilingual collaborative ontology development in wikidata," in Companion of the The Web Confer- ence 2018 on The Web Conference, 2018, pp. 23-27

LONG-TERM PRESERVATION OF A SOFTWARE EXECUTION STATE

Rafael Gieschke
University of Freiburg
Germany
rafael.gieschke@rz.uni-freiburg.de
0000-0002-2778-4218

Klaus Rechert
University of Applied Sciences Kehl
Germany
rechert@hs-kehl.de
0000-0002-2454-4374

Euan Cochrane
Yale University Library
U.S.
euan.cochrane@yale.edu
0000-0001-9772-9743

Abstract – Software is a very complex product, offering an endless number of different states and appearances. To foster academic discussion about software-based cultural and scientific phenomena like computer games, digital art, or scientific computational models, it is necessary to be able to reference specific moments of running software. In this article, we discuss the possibilities to "freeze" software while being executed and describe constraints for the long-term preservation of these snapshots.

Keywords – Emulation, Program Snapshots

Conference Topics – Immersive Information

I. INTRODUCTION

Citation is an important tool for scientific research. With citations, one can point to prior or related work, for comparison and scientific or socio-cultural discussion. Hence, access to cited work is an essential service of libraries and archives. Long-term digital preservation is the tool to ensure the availability of (cited) digital works for future access.

With the emergence of more and more complex born-digital works, such as software-based art, computer games, and all kinds of interactive digital artifacts, and software in general, citation of works as well as access to cited works became a significant challenge. Citation of software has been integrated into scholarly practice, firstly due to the increased importance of software as a tool for research, e.g., data processing and data modeling. Especially since scientific software is usually made for a specific purpose, requires dedicated resources, and is an indispensable part of the research process and, thus, their authors deserve attribution. Furthermore, software setups perform extensive and complex operations. With the requirement to reproduce scientific results, availability and ideally re-executability of software is curial.

With emulation, there is a rather generic approach to keep software artifacts usable [1]. Providing a usable (or executable) reference to access and use software artifacts is currently ongoing work, e.g., through the Emulation as a Service Infrastructure (EaaSI) program of work [2]. A working and usable infrastructure to (re-)execute or "re-perform" a preserved digital object is necessary but not always sufficient to support research activities. Software or digital objects in general are not only complex in technical terms and, thus, need necessary technical infrastructure to be rendered but also complex in use and operation and manifold in their options, appearance, and how they are perceived by users. Hence, in many cases the description of the software and its execution dependencies is not sufficient to describe specific aspects of the software's performance. The basic assumption of software functionality (and computers) to process data input and (re-)produce an output does not cover intermediate steps or user-machine and machine-machine interactions during execution. These, however, do not only bear important aspects of how a particular result has been achieved but may constitute important information or facts themselves. For instance, a computer game or a software-based digital artwork can produce an almost endless combination of states (e.g., the state after a character in the game has been positioned in a particular way after responding to other game elements). Not all of them are relevant, but some are. Picking these for

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

scholarly discussion requires a way for future readers to explore these states to contribute to the scientific discourse.

In general, one can imagine different options providing future scholars access to these states. The first and technically most viable as well as possibly the most cost-effective option is to create (static) documentation, e.g., through screenshots or video captures [3]. However, this method limits future research options, e.g., it will be difficult for future researchers to explore and inspect the context as well as possible follow-up states.

If additionally, the software and its execution dependencies are preserved and accessible, the documentation can be used to perform necessary steps manually to reproduce a certain state or result. In most cases, however, a manual "replay" of documented steps can be difficult and time consuming. The documentation may not always be sufficient to guide future users successfully through the process, since users may lack skills and/or implicit operational knowledge. Especially computer games require some training to perform certain actions successfully. Furthermore, it is still not assured (especially in the case of a manual re-play) that the state reached is identical to the one described in literature. A potential technical solution to this problem is capturing user-interactions and support automated deterministic replay. Even though this approach shows promising results, there are still a lot of difficulties to be solved for generic use-cases, due to non-deterministic events or behavior of the system, especially with highly dynamic interactions [4].

A further option is the possibility to "freeze" the execution of software at any chosen state as an execution snapshot. Ideally, such snapshots could be archived, shared with others, and restored at any time to continue the execution of the software. While creating snapshots of running virtual machines (VMs) is a common feature of most virtualization and many emulation systems, reactivation in (slightly) differing technical setups as well as long-term usability of snapshots is an open issue. In this article, we investigate this problem and describe a way to implement long-term available "frozen" snapshots of running software.

1. Related Work

With the growing importance of software as a cultural and scientific product, the necessity of citing software was recognized [5]. Citation practice ranges from informal mention to detailed version information but are being formalized further. Citation of software implies (or should imply) the availability of software.

Even though many initiatives have started archiving software, software preservation in general is still an open issue because of the complexity and size of the task and sometimes due to legal obstacles. Preservation of source code [6] is to a certain degree the most systematic and successful public approach so far. However, available source code does only cover a limited field (open-source software) and, more importantly, there is a long-term usability problem [7], as source code cannot be trivially reused without further technical steps (i.e., compilation into an executable binary form). Furthermore, for more complex scenarios, different software products need to be contextualized (setup and configured) within a defined execution environment (e.g., using an emulator) and, especially in context of computational research, brought together with data. Container technology solved some of the dependency, setup and configuration problems [8] but seldomly reduces the operational complexity (e.g., the number of settings, runtime variables) of non-trivial setups.

For smaller, controlled subsets, there are tailored solutions, not only to reproduce a computational result but also to transparently document all in-between steps in an executable form, e.g., Jupyter notebooks. But even for these rather constrained niches, long-term access and reproduction is quite difficult [9], [10].

Another alternative to deterministically restore a specific application state without reproducing interactive user (GUI-)events (e.g., [11]), which are prone to non-deterministic behavior of GUI applications, are so called "record and replay" tools (e.g., [12], [13]). In contrast to tools recording GUI-events (e.g., macro recorders), these tools record the interaction of an application with the underlying operating system and, thus, can deal with potential non-determinism (e.g., clock/time functions, random number generators). These tools, however, are usually quite intrusive (e.g., they depend on special CPU features or only work for special cases like

applications with limited interactivity) and slow down the execution of the recorded application significantly.

2. Snapshots of Running Application Processes

Presumably, the main motivation to access preserved (and cited) software setups beside verification of published results could be exploring the relevant software environment further, e.g., by changing intermediate inputs in a case of interactive computational science or to follow different paths of a computer game. But there are many other different use-cases for preserving a specific state of a running software setup. This section will discuss the concepts to generate software snapshots.

A "snapshot" is a saved (serialized) state of a running process or system. Any information needed to restore the process or system and to continue its execution is contained within the snapshot. This operation requires some technical support of the surrounding environment, i.e., the software that is creating the snapshot. Since we do not want to constrain snapshots to a single architecture, computer platform, or operating system, i.e., we want to be able to cover applications from different technical epochs, we chose to take a snapshot of the whole system the application of interest is running in (guest system). We, thus, assume that the guest system is running on either virtualized or emulated hardware (i.e., in an emulator or hypervisor).

In principle, there are different ways to create a snapshot of a running system. The first one is taking advantage of the capabilities of the hypervisor or emulator. Some emulators (e.g., QEMU, v86) or VM hypervisors (e.g., VMware Workstation/ESXi, VirtualBox) offer built-in functionality to create snapshots. The emulator or hypervisor can pause the execution of a guest and save all states of the hardware (i.e., the virtual CPU and any other virtual hardware devices). The snapshots created this way are called *virtual machine snapshots*. Saving the state of the whole computer system to disk can also be done by the (guest) operating system itself without the help of a hypervisor, a feature typically called "hibernation" or "suspend to disk". However, this option is only available for a few emulators/hypervisors (and guest operating systems in case of hibernation).

But having a working option to create virtual machine snapshots (e.g., QEMU), a further option to consider could be to run emulators not capable of creating snapshots themselves (e.g., SheepShaver, VICE) within a capable hypervisor (or emulator) and use it (the "outer" emulator) to create snapshots (of the "inner" emulator). Most likely, the generated snapshots depend on the software (emulator or hypervisor) they were created on. Firstly, virtual machine snapshots are saved (serialized) in a format specific to the emulator/hypervisor software used. Secondly, while it would be possible for other emulators to re-implement support for an existing virtual machine snapshot format, this seems unlikely as the virtual machine snapshots including their configured operating system require the exact same (virtual) hardware (i.e., not only the same CPU but also identical other virtual hardware devices) it has been created on to be restored successfully [14]. Even the forward-compatibility of virtual machine snapshots to later versions of the same emulator/hypervisor will likely decrease over time as their main usage is for live migration and short-term snapshots.[1] Hence, virtual machine snapshots must be carefully curated and maintained together with the necessary software (i.e., the emulator or hypervisor) for restoring, which will lead to significant maintenance overhead over time.

An alternative option to storing the complete hardware state of a virtual machine is to create *process snapshots* (also known as application snapshots) which do not rely on a hypervisor but on functionality offered by the operating system. Process snapshots save the state of a specific running application, comparable to the way the operating system saves the process's state when executing a context switch.

[1] We have found that a snapshot taken using hibernation by the guest operating system might even fail to restore on different builds of the very same emulator source code, see https://gitlab.com/emulation-as-a-service/emulators/qemu-eaas/-/blob/v2.5/dsdt.patch. This finding might be extendable to virtual machine snapshots taken by the emulator/hypervisor but does require further research.

While virtual machine snapshots save the state of the whole platform (i.e., the CPU and any devices, e.g., the IBM PC platform), process snapshots only save the state of (parts of) the CPU (i.e., only the parts of the hardware defined by the instruction set architecture (ISA), e.g., the x86-64 architecture). For other resources (the equivalent of devices), they make use of the abstraction the operating system provides and are therefore less hardware dependent. Consequently, however, they have a rather strong operating-system (i.e., kernel Application Binary Interface "ABI") dependency as well as a dependency on any open files, network connections, or similar resources originally being in use by the process at the time of creating the snapshot. For our purposes, we chose to create a process snapshot of the emulator process[2], and thus, implicitly the state of the guest system.

Taking a snapshot of a running process usually requires special support from the operating system, because some parts of the process's states are not visible to the process itself but only available to the operating system. The Linux kernel was extended to provide previously missing APIs to create process snapshots by the *Checkpoint/Restore In Userspace* (CRIU) project[3], which also provides a user-space helper utilizing the kernel's API to create, serialize (i.e., save as regular files), and restore process snapshots[4].

CRIU works best when it is used to snapshot isolated applications running in their own Linux container[5] using a container runtime[6] and is, in fact, already well integrated into contemporary container runtimes[7]. As we already developed a framework to package (untrusted and potentially insecure) emulators including supporting infrastructure as self-contained container images and let them run in an isolated Linux container in the EaaS emulation framework [15], we used the same approach for CRIU[8]. Taking this approach, preserving emulators (and snapshots) becomes a special case of preserving containers [16].

3. Long-Term Access to Process Snapshots

Solving portability issues of the created snapshots solves only the smaller part of the problem. The second problem set is to restore or reactivate a snapshot using future computer systems. By choosing process snapshots, we have reduced hardware dependencies, which are typically abstracted by the operating system. For instance, for a (user-space) application to be run on a Windows or Linux system, the hardware configuration of the computer does typically not matter to the application. Instead, an application uses well defined interfaces of the operating system to interact with graphics, sound, or network hardware. These interfaces are stable and do not change, e.g., if a hardware component is replaced. Hence, for process snapshots, the remaining (future) dependencies are the operating system application binary interface (ABI) and the CPU instruction set architecture (ISA; e.g., x86-64), which should make these snapshots portable between contemporary systems but also to future systems.

---

[2] We are only using emulators here. There would be no principal problem in creating a process snapshot of a hypervisor that uses the Linux kernel's KVM ABI (e.g., QEMU) as these hypervisors are normal processes. However, this functionality has only niche use cases and is not supported by process snapshot frameworks, see https://github.com/checkpoint-restore/criu/issues/229.

[3] Checkpoint/Restore In Userspace, see https://criu.org/.

[4] The kernel patches created by the CRIU project were accepted and included in the upstream Linux kernel, so that the CRIU user-space helper can be used together with any stock Linux kernel, reducing maintenance burden.

[5] See, e.g., https://www.redhat.com/en/topics/containers/whats-a-linux-container.

[6] Standardized in the OCI Runtime Specification (https://github.com/opencontainers/runtime-spec) and implemented in, e.g., Docker, runc, or crun with the help of functionality provided by the Linux kernel.

[7] See, e.g., https://github.com/opencontainers/runc/blob/main/man/runc-checkpoint.8.md.

[8] The images may have to be slightly altered (e.g., checked in an automated process) as CRIU can place subtle restrictions on the kind of executables it is able to restore properly, see https://github.com/checkpoint-restore/criu/issues/1507.

1. *Restoring the Execution Context*

When CRIU restores a snapshot, it interacts with the Linux kernel to restore the exact same state the process was in when the snapshot was created. The "exact same state" reproduced by CRIU includes CPU registers and memory used by the application as well as any resources opened by the application, e.g., files or network sockets. These resources must remain available in the very same state (e.g., the same file content) and must be saved, managed, and restored independently from the CRIU snapshot. We have already accomplished this by using container images, for which a derivative image can be created when snapshotting the application. For network connections (e.g., connected TCP sockets), CRIU's approach is usually more brittle as they depend on the uncontrollable outside world (e.g., the remote end of a TCP socket will probably be gone when the snapshot is restored). In the EaaS emulation framework, we have already solved this problem by providing network access only via a virtualized and isolated network.

2. *Identifying Remaining Hardware Dependencies*

The described approach makes sure that any applications can continue to run after being restored. However, CRIU's intended usage scenarios[9] are focused on short-term usage of the created snapshots on homogenous machines (i.e., using the same CPU model or, at least, the same CPU generation from the same CPU vendor with the same architectural features, similar other installed hardware devices, and similar Linux kernel versions). In other words, CRIU's process snapshots may still depend on the original CPU model. This can be a surprising property as applications can break in unexpected and subtle ways under different CPU models, even more so when the CPU is changed during the application's execution. Normal applications do not expect that the CPU's (visible) features change during execution[10] and, as the application, deliberately, has no way to notice that it is being snapshotted and restored, it even has no chance to react to the changed CPU at all.

On the x86(-64) architecture, the CPU vendor and model, its supported extensions to the original ISA, and other features of the CPU are exposed via the CPUID machine instruction. The CPUID instruction is unprivileged and, thus, can be executed by user-space applications. It is typically, transparently, included by the compiler into the executable binary[11] for features like function multiversioning[12] or queried using dedicated libraries[13] by the application itself. All these features have in common that they, at the start of the application, select one of several machine code versions of the same function most suitable (i.e., optimized) for the current CPU model. This is a problem if the target CPU has less features (e.g., no AVX2) than the original CPU. After resuming the snapshot, the application will still execute the code path requiring the original CPU features (e.g., AVX2) and execution on the target CPU will fail, generating a SIGILL signal (on Linux/POSIX) and terminating the process[14].

A possible approach could be to require the target CPU to always support more features than the original CPU, e.g., by requiring it to support all CPU extensions available at a time. We found, though, that this approach does not work as, apart from being expensive by requiring the latest CPUs, CPU features are not only added by vendors but sometimes also removed again in later CPU model

[9] https://criu.org/Usage_scenarios

[10] For their recent processors with heterogeneous CPU core configurations, Intel takes great pain in ensuring that all cores expose exactly the same process-visible features, see, e.g., https://www.tomshardware.com/news/intel-nukes-alder-lake-avx-512-now-fuses-it-off-in-silicon.

[11] See, e.g., gcc and its libgcc: https://github.com/gcc-mirror/gcc/blob/59a72acbccf4c81a04b4d09760fc8b16992de106/gcc/common/config/i386/cpuinfo.h#L975.

[12] https://gcc.gnu.org/onlinedocs/gcc/Function-Multiversioning.html

[13] See, e.g., libcpuid: https://github.com/anrieff/libcpuid.

[14] While it would be imageable to catch this signal and emulate just the missing (e.g., AVX2) instructions, in practice, this approach is unfeasible as other (existing) (SSE) instructions modify the state of the (extended) registers used by AVX2 as well while not producing a catchable SIGILL instruction. It is, thus, not possible to selectively emulate these instructions but all instructions (including the existing ones) would have to be emulated.

generations[15]. A "best" CPU including every feature ever introduced may, thus, not always exist.

More importantly, though, we found that, e.g., a snapshot created on a CPU not supporting AVX2 will crash on a CPU supporting AVX2, i.e., a CPU with strictly more features. Debugging showed that this is due to the XSAVE instruction, which writes the CPU register's state to memory at an application-provided location. The (byte) size the register state will need in memory can be queried by the application (or a supporting library) using the CPUID instruction, but this, again, is typically only done at application startup and cached for later use. If, in our case, the snapshot is now resumed on a target CPU requiring more space for its register state (i.e., by having the larger AVX2 registers), not enough space will have (unknowingly) been reserved by the application when calling the XSAVE instruction and the CPU will silently overwrite parts of memory, e.g., in the application's stack, leading to the application's crash.

A possible workaround is to disable the problematic features (e.g., AVX2) on the target CPU with the operating system's help[16]. This workaround was tested but proved not to be successful as a number of applications, including the widely used GnuTLS library[17], check incorrectly for the availability of CPU features, subsequently still try to use the disabled features and crash. This results in crashes of even very basic dependent applications like "apt-get".

Thus, the most feasible way is to directly restrict the CPUID instruction to report only desirable features as available and report a buffer size for the XSAVE instruction that is large enough for any target CPU's register state. At the same time, this is advantageous as it allows to restrict the used CPU features to a sensible set of features supported by not only the latest CPUs but a large number of (cheaply available) CPUs from different vendors (and most x86-64 emulators). Such "common denominators" of features are already standardized as micro-architecture levels (e.g., x86-64-v2) in the ELF x86-64 psABI[18] and are recently starting to be used to define minimum system requirements of Linux distributions.

CPUID virtualization

Manipulating the CPUID instruction, though, proves to be problematic on the x86-64 architecture, a property than can be attributed to the fact that the x86-64 architecture (without extensions) does not conform to the virtualization requirements introduced by Popek and Goldberg [17]. The CPUID instruction can be seen as a behavior sensitive instruction that is not privileged, i.e., can be executed directly by user-space application without a chance for the operating system's kernel to interfere.

An extension found on most Intel processors[19] remediates this problem by allowing to turn the CPUID instruction into a privileged instruction, and, thus, trapping it in user space.[20] This feature is used

[15] Again, see the removal of AVX-512 from Intel's recent processors with heterogeneous CPU core configurations, https://www.tomshardware.com/news/intel-nukes-alder-lake-avx-512-now-fuses-it-off-in-silicon.

[16] Using the Linux kernel's "clearcpuid" and "noxsave" command-line options, which do not directly interfere with the result of the CPUID instruction for user-space applications but only disable CPU features using the CR4 control register.

[17] See https://gitlab.com/gnutls/gnutls/-/issues/1282. The underlying problem is that such problems do not get much real-world test coverage as users do not typically want to restrict their CPU's features, i.e., never run applications under such kernel configurations.

[18] Processor-specific application binary interface, https://gitlab.com/x86-psABIs/x86-64-ABI.

[19] Exposed by the Linux kernel via the arch_prctl(ARCH_SET_CPUID, ...) system call, see https://man7.org/linux/man-pages/man2/arch_prctl.2.html.

[20] Both Intel's VT-x and AMD's AMD-V virtualization extensions for the x86-64 architecture do allow for hypervisors to trap CPUID instructions. An alternative approach would, thus, be to have the application run in its own virtual machine (potentially inside yet another virtual machine provided by a cloud provider) restricted to, e.g., the x86-64-v2 micro-architecture level. This approach was not pursued as more virtualization levels will most

by the libvirtcpuid project[21], which, independently from our work, researched the presented problem of a snapshot created on a CPU with less features (e.g., no AVX2) crashing when being resumed on a CPU with more features (e.g., AVX2). Their intended use case, however, focuses on live migration of applications and differs from our long-term preservation use case. Additionally, they rely on the described Intel extension, which is neither found on AMD x86-64 CPUs nor in emulators and often not exposed by cloud providers, reducing its usefulness for our application significantly.

We, thus, modified libvirtcpuid's approach slightly: while the original libvirtcpuid relies on the described CPU extension to trap CPUID instructions, we modify the application's executable (ELF) binary files in advance to replace every CPUID instruction with an RDMSR instruction[22]. RDMSR instructions, in turn, are always privileged on the x86-64 architecture and, thus, will trap with a signal to user-space that can be processed by libvirtcpuid in its usual way[23]. This approach is not guaranteed to work as applications may dynamically generate just-in-time (JIT) code including CPUID instructions. Only being generated at runtime, these instructions would not be processed by our tool and still leak unmodified CPUID information to the application. However, this is very unlikely as, as described above, code using the CPUID is typically statically generated by the compiler or placed in dedicated libraries. Other potential problems, e.g., applications checking for CPU features by trying to directly use them without checking for their availability first[24], are the same as for the upstream libvirtcpuid and rare in practice.

Other sources of non-determinism

As described above, the restored application also depends on outside resources. Obvious ones like files (either regular files, UNIX domain sockets, or pipes) are already handled by our emulation framework. Non-obvious ones include time functions[25], which naturally depend on the time in the outside world, either measured as time since the system was booted (CLOCK_MONOTONIC) or as real (i.e., wall-clock) time (CLOCK_REALTIME). The former (CLOCK_MONOTONIC) can be virtualized by the Linux kernel using time namespaces[26], which are already supported and handled by CRIU. As they are typically used via the C standard library, the latter (CLOCK_REALTIME) can potentially be modified in user-space[27] to be derived from CLOCK_MONOTONIC with a constant (user-configurable) offset.

However, an application could also use the RDTSC instruction to directly read the processor's time-stamp counter. Differently from the CPUID instruction, the RDTSC instruction can be disabled

probably degrade performance and (nested) virtualization is not universally available at cloud providers or comes with extra costs, see, e.g., https://ignite.readthedocs.io/en/stable/cloudprovider/.

[21] https://github.com/twosigma/libvirtcpuid

[22] Introducing the ELFant tool, a powerful but friendly shell script that tramples over your ELF files, see https://github.com/emulation-as-a-service/libvirtcpuid.

[23] A conceivable alternative approach of replacing CPUID instructions with a call to a library function emulating and manipulating the CPUID instruction directly is not feasible as the CPUID instruction is encoded in only 2 bytes, which is not sufficient space for any jump/call instruction. In contrast, the RDMSR instruction is encoded in 2 bytes as well, and can, thus, directly replace the CPUID instruction in the binary file without any further modifications. Yet another conceivable alternative approach of replacing library functions (e.g., libcpuid) utilizing the CPUID instruction is not feasible as they might not easily be recognized anymore when linked into the application in binary form, the variety of such libraries is too diverse, and some applications (e.g., GnuTLS) do not employ such libraries at all but directly use the CPUID instruction for similar purposes, leading to a very brittle application-specific manual patching approach.

[24] In this case, if the CPU feature is available on the original CPU, the application will recognize it as usable but will crash as soon as trying to use it on the target CPU after the snapshot is resumed.

[25] Exposed via the clock_gettime() system call on Linux, see https://man7.org/linux/man-pages/man3/clock_gettime.3.html.

[26] https://man7.org/linux/man-pages/man7/time_namespaces.7.html

[27] For instance, using the LD_PRELOAD mechanism, see https://man7.org/linux/man-pages/man8/ld.so.8.html.

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

and trapped in user-space[28]. As the application might not expect RDTSC to not work, further modifications to the application might be necessary. In contrast to the CPUID problem, however, this problem is immediately visible already when starting the application on the original CPU and can, thus, be handled more easily instead of manifesting itself in an unfixable way later long after the snapshot has been created.

A final class of resources that must be dealt with are resources available on the original system that are being masked by the container runtime/configuration but are not available at all on the new system. As they are not available at all, CRIU cannot mask them like on the original system and, thus, might fail to restore the snapshot.[29] Here, a simple workaround is to not mask the resources in the first place.[30]

5. *Emulation in Emulation*

To restore an application state, the snapshotted emulator process containing the running application must be restored using a suitable technical environment. In the case of a future reactivation, the required technical environment for restoring will be an appropriate emulator, e.g., a suitable x86-64 emulator satisfying the CPU dependencies (e.g., the x86-64-v2 micro-architecture level) and a container runtime including an appropriate Linux kernel and pre-configured tools to restore the snapshot (e.g., CRIU). This runtime then will be able to restore the snapshotted process (emulator running a guest system). This setup, however, will lead to an emulation-in-emulation (stacked emulation) access context.

Using emulation-in-emulation as an emulation-based digital preservation strategy is not an ideal solution in general. While the idea is simple and appealing, i.e., today's emulator setups containing obsolete systems and running, e.g., on a contemporary Windows 11 system, can be preserved and kept available through future emulators by simply focusing on today's Windows 11 system and so on. However, with technical epoch and thus, every new level in this emulator stack, a technical and conceptual mapping between contemporary computer systems and the last generation (latest emulators) must be made.

These mappings usually require technical compromises, especially but not only, for interactive usage, because future concepts have changed significantly. For instance, by moving toward gesture inputs, mapping modern touchscreen gestures to 3-button mouse events is necessary. Clever emulator developers will find a user-friendly and usable solution for their technical environment and context. However, with every additional layer (and mapping), it will be harder to operate the original system (e.g., the 3-button mouse, through an emulated touchscreen using a future VR setup) and, eventually, some states or concepts of the original system will become inaccessible throughout the different layers. Therefore, we usually argue to "migrate" the old (guest) systems to new emulators, such that any new generation of computer systems with new interaction paradigms (e.g., virtual reality) adapt these directly to the old concepts.

For this special case (snapshots of running processes), however, the emulation-in-emulation scenario is necessary and justifiable. Its necessity results from the design choices of taking process snapshots. In the case of virtual machine snapshots, restoring the saved machine state on future emulators is possible in theory. However, in practice it will be quite difficult since the target machine must match exactly the hardware of the snapshotted virtual machine. Even the transition (live migration) between two contemporary virtual machines, both running within the same emulator (in this case, two

[28] Exposed as prctl(PR_SET_TSC, …) by the Linux kernel, see https://man7.org/linux/man-pages/man2/prctl.2.html.

[29] For instance, the Linux kernel only provides /proc/asound/ (masked as empty directory inside containers) if the host system includes a sound card. This is typically true for desktop computers but untrue for server or cloud computers. A snapshot created on a desktop computer may, thus, fail to resume on a server or cloud computer.

[30] This can pose security problems but, e.g., in the sound card scenario, it might be acceptable to not mask the sound card when being run on a desktop computer as the desktop computer is typically operated by only one user, who is already able to manipulate and interfere with the sound card in host system anyway.

QEMU-based VMs) turned out to be rather difficult [18].

For the conceptual idea of saving and restoring an application state, the concept of emulation-in-emulation is appropriate. Not only is the depth of the emulation stack limited to the maximum of one extra layer and, thus, the mapping problem between contemporary systems and old systems is addressable, but this also is a desired setting since the future user is able to observe the exact state, including all features and limitations of the access platform (emulator) the creator of the snaphot has experienced. Even if the restored snapshot offers limited usablity (compared to running the application using future emulation-based access platforms), it offers a stable and reproducible reference point.

4. Conclusion and Future Work

In this article, we have presented a technical and conceptual analysis on the preservation of software execution states (so called snapshots). The concept of preserving snapshots will not only contribute a further facet for software citation, but it will also contribute to an increased usability of emulation setups by simplifying the preparation of ready to use software setups. Users can be presented with a configured and running application in a usable state, without the hurdle of operating an old computer system, e.g., staring an application, finding the necessary files, etc. For some (future) software setups, such an approach might be the only viable solution. Software becomes a boundless product which is difficult to capture and to "own" since it is not shipped anymore on media. Modern software offers dynamic installation processes with multiple options for extension, in-app purchases, etc., relying not only on individual decisions but also the publisher's infrastructure and especially support of installing outdated software packages.

In addition to citation use cases, the functionality required to fulfill these can be useful for other long-term access purposes. For example, when making digital artifacts available via emulation that rely on slow or complex software environments, it can be valuable from a user-experience perspective to be able to immediately restore the system or network state to a point after all the software components involved have completely loaded. The state-saving functionality described in this paper is currently in use at Yale University Library as part of their efforts to retain access to web sites including the Ross Archive site[31] and the Historical Register Online site[32]. Access to both sites is provided via emulation of hardware supporting the underlying web server software, database software, and a contemporaneous web browser. Each of these has to be loaded before the user can access the site in emulation and the process can take minutes. By pre-loading them, saving the state in a process snapshot, and loading the state at point of access, this saves the user a great deal of time and significantly improves the user experience for them.

While there is a proof-of-concept implementation available as well as working real-world examples as part of the EaaSI project, the focus of this work was to improve our understanding of potential technical hurdles restoring snapshots in the future. x86-64-based systems are still the most important platform running in-production framework. Hence, our analysis was focused on x86-64 process snapshots. Future work will widen the scope to other relevant platforms (e.g., ARM64).

Re-storing snapshots does not only pose technical challenges but also administrative ones. The runtime environment must be archived and maintained, as well as any other runtime dependency. By encapsulating emulators with their runtime dependencies within containers, preserving containers (and their runtime), the preservation of snapshots is just a special case of the existing EaaS container preservation workflow.

An important limitation of our experimental setup is the time lag between a snapshot request and the snapshot executions. This lag is currently somewhere around 2-5 seconds. Additionally, serializing a snapshot takes time – linear to the total memory of the process used. Hence, currently our setup is not yet suitable for highly dynamic software setups, e.g., computer games, where states are changing very quickly, and not suitable to create a series of fine-grained snapshots. However, there are promising developments to improve snapshot

[31] https://rossarchive.library.yale.edu/

[32] https://yalehistoricalregister.library.yale.edu/

performance and to support fine-grained, high frequency incremental snapshots [19].

5. REFERENCES

[1] D. S. Rosenthal, "Emulation & virtualization as preservation strategies," 2015.

[2] E. Cochrane, K. Rechert, J. Oberhauser, S. Anderson, C. Fox, and E. Gates, "Useable Software Forever," *IPres 2022 Glasg. 12—16 Sept. 2022 Www Ipres2022 Scot*.

[3] J. P. McDonough *et al.*, "Preserving virtual worlds final report," 2010.

[4] J. Oberhauser, R. Gieschke, and K. Rechert, "Automation is Documentation: Functional Documentation of Human-Machine Interaction for Future Software Reuse," *Int. J. Digit. Curation*, vol. 17, no. 1, Art. no. 1, Sep. 2022, doi: 10.2218/ijdc.v17i1.836.

[5] A. M. Smith, D. S. Katz, and K. E. Niemeyer, "Software citation principles," *PeerJ Comput. Sci.*, vol. 2, p. e86, 2016.

[6] R. Di Cosmo, M. Gruenpeter, and S. Zacchiroli, "Referencing source code artifacts: a separate concern in software citation," *Comput. Sci. Eng.*, vol. 22, no. 2, pp. 33–43, 2019.

[7] K. Rechert, J. Oberhauser, and R. Gieschke, "How Long Can We Build It? Ensuring Usability of a Scientific Code Base," *Int. J. Digit. Curation*, vol. 16, no. 1, p. 11, 2021, doi: 10.2218/ijdc.v16i1.770.

[8] C. Boettiger, "An introduction to Docker for reproducible research," *ACM SIGOPS Oper. Syst. Rev.*, vol. 49, no. 1, pp. 71–79, 2015.

[9] J. F. Pimentel, L. Murta, V. Braganholo, and J. Freire, "A large-scale study about quality and reproducibility of jupyter notebooks," in *2019 IEEE/ACM 16th international conference on mining software repositories (MSR)*, 2019, pp. 507–517.

[10] J. Wang, T. Kuo, L. Li, and A. Zeller, "Assessing and restoring reproducibility of Jupyter notebooks," in *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering*, 2020, pp. 138–149.

[11] T.-H. Chang, T. Yeh, and R. C. Miller, "GUI testing using computer vision," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 1535–1544.

[12] O. S. Navarro Leija *et al.*, "Reproducible containers," in *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, 2020, pp. 167–182.

[13] R. O'Callahan, C. Jones, N. Froyd, K. Huey, A. Noll, and N. Partush, "Engineering Record and Replay for Deployability.," in *USENIX Annual Technical Conference*, 2017, pp. 377–389.

[14] K. Rechert, O. Stobbe, O. Zharkow, R. Gieschke, and D. Wehrle, "CiTAR - Preserving Software-based Research," *Int. J. Digit. Curation*, vol. 15, no. 1, Art. no. 1, 2020, doi: 10.2218/ijdc.v15i1.716.

[15] R. Gieschke and K. Rechert, "A Generic Emulator Interface for Digital Preservation," in *iPres 2022 Glasgow 12—16 September 2022 www. ipres2022. scot*, 2022.

[16] K. Rechert, T. Liebetraut, D. Wehrle, and E. Cochrane, "Preserving containers–requirements and a todo-list," in *Digital Libraries: Knowledge, Information, and Data in an Open Access Society: 18th International Conference on Asia-Pacific Digital Libraries, ICADL 2016, Tsukuba, Japan, December 7–9, 2016, Proceedings 18*, 2016, pp. 225–230.

[17] G. J. Popek and R. P. Goldberg, "Formal requirements for virtualizable third generation architectures," *Commun. ACM*, vol. 17, no. 7, pp. 412–421, 1974.

[18] J. Wei, L. K. Yan, and M. A. Hakim, "Mose: Live migration based on-the-fly software emulation," in *Proceedings of the 31st Annual Computer Security Applications Conference*, 2015, pp. 221–230.

[19] R. S. Venkatesh, T. Smejkal, D. S. Milojicic, and A. Gavrilovska, "Fast in-memory CRIU for docker containers," in *Proceedings of the International Symposium on Memory Systems*, 2019, pp. 53–65.

RETHINKING DIGITAL PRESERVATION

Conceptual Foundations

Stephen Abrams
Harvard University
United States
stephen.abrams@harvard.edu
ORCiD 0000-0003-2326-6672

Abstract – In support of a multi-year initiative to revitalize its core digital preservation infrastructure, the Harvard Library is engaged in an open-ended exploration of an ideal system solution. The individual components of that ideal cohere into abstract functional and informational reference models, which act as aspirational benchmarks for requirements and subsequent procurement and deployment activities. The models are derived through the logical refinement of a small set of high-level axiomatic principles. These reflect a conceptualization of digital preservation as an inherently communicative enterprise with an ultimate goal of complementing the persistence of authentic digital information objects with that of opportunities for legitimate information experiences.

Keywords – abductive inference, abstract reference model, communicology, finite state machine, information experience

Conference Topics – From Theory to Practice

I. INTRODUCTION

The Harvard Library began operation of its Digital Repository Service (DRS) in October 2000. At that time, no viable commercial or open-source products were available. Consequently, it was necessary for the Library to build a novel system in-house [1]. Since then, use of the DRS has grown to hosting over 10.6 million digital objects, 890 million files, and 90 formats, totaling over 2 PB. These materials span all content genres critical to the University's research, teaching, and learning mission as well as its administrative operation. While the DRS technical platform has been maintained and incrementally updated over the past two decades [2][3], it still remains a custom system making increasingly unsupportable demands on finite internal resources. Furthermore, the functional applicability of the DRS is increasingly constrained by limitations arising from long-standing and deep-seated conceptual design, implementation, and operational decisions. To address these concerns, the Library is engaged in a generational modernization known as the DRS Futures project. This effort will revitalize the DRS and reposition it to continue to provide effective, efficient, and sustainable stewardship of the University's digital collections in light of future challenges and opportunities [4].

The Futures project is structured in three phases:

1. Envisioning an *ideal* repository
2. Specifying an *achievable* repository
3. Deploying an *operational* repository

The first phase is a purposefully open-ended investigation of aspirational needs and goals explicitly unfettered by considerations of how they ultimately will be provisioned. These ideals will be winnowed down to the achievable in the second phase, contextualized with the aspirational end-goals foremost in mind. In essence, the Library is looking beyond what the state-of-the-art might be today, towards what it could and should be in the near or far future. Such long-range strategic thinking is possible only when rooted in robust philosophical and conceptual foundations.

1. EXPLORATORY APPROACH

The process of planning and deploying any significant socio-technical system naturally progresses through stages of initial ideation and subsequent development or procurement [5]. The transition from the intangible considerations of the former to the specifics of the latter is codified in terms of system requirements. These function variously as a specification for development, an

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

evaluative rubric for procurement, and acceptance criteria for formal project completion. Traditional requirements development is approached *inductively* [6], relying on stakeholder engagement as well as reference to prior practice, professional intuition, and shared community attitudes to establish needs, goals, and aspirations ultimately refined into a set of use cases [7]. However, in order to achieve a higher level of confidence in final requirements, inductive results should be complemented by a parallel *abductive* process deriving requirements from a small axiomatic set of accepted first principles [8].

Andow describes abduction as the mode of logical inference that seeks the best possible explanation for a domain's phenomena, in distinction to deduction's logically-necessary and induction's logically-most-probable explanations [9]. The final logical refinement of these philosophical and conceptual principles constitutes an abstract reference model (ARM) of the desired system. An ARM is a framework defining the fundamental entities and relationships constituting a domain untethered from the semantics of any specific implementations [10].

Due to its logical formality and systematic application, abductive derivation is more likely to result in comprehensive coverage of appropriate domain considerations relative to a more ad hoc and anecdotal inductive process, however well-grounded it may be in historical precedent, domain best practice, and professional experience. In essence, the top-down abductive approach starts with a high-level model of the entire domain under consideration and systematically segments it into smaller and smaller units of greater and greater conceptual detail. The bottom-up inductive approach, on the other hand, starts with various granular units of detail that are gradually refined and abstracted with an assumption that they will eventually cohere into comprehensive coverage of the full domain. Ideally, the two approaches will exhibit significant, if not full, overlap. Regardless, performing the two activities in parallel provides an opportunity to identify and fill in any gaps resulting from the individual exercises.

2. Philosophical Inquiry

The foundational basis for the Futures project emerged through a process of Philosophical Inquiry (PI). PI is a qualitative research method deriving meaning from experience through abductive questioning of fundamental assumptions within a domain of practice to propose new, and better, explanatory structures for that domain [11]. In the Futures context, the inquiry began with questions regarding the fundamental nature of the preservation enterprise. The *Encyclopedia of Archival Science* defines digital preservation as "the processes and controls that enable digital objects to survive over time" [12]. This formulation emphasizes an object- and process-centric view that implicitly promotes a metaphoric narrative of digital preservation as a *managerial* endeavor. That is, a set of activities done *to* objects to ensure persistence of their significant characteristics over time. While an important foundational step, this narrative minimizes critical attention to what subsequently can be done *with* those objects and to what *effect*.

Similarly, the phraseology common to other community-accepted definitions of the preservation field – for example (with emphasis added): "policies, strategies, and *actions* that ensure *access* to digital content over time" [13]; "act of *maintaining* information, independently Understandable by a Designated Community, and with evidence supporting its Authenticity, over the Long Term" [14]; "series of *managed* activities necessary to ensure continued *access* to digital materials for as long as necessary" [15]; "*processes* aimed at ensuring the continued *accessibility* of digital materials" [16] – emphasizes two points. First, that the primary role of domain agents is an *enabling* one, e.g., acting as strategi*zers*, maintain*ers*, manag*ers*, process*ers*. Second, that the imperative goal of the exercise is provision of artifactual access.

Access refers to the ability and permission to find and retrieve information relevant for a specific purpose [17]. In other words, access is an enabling factor for subsequent use, which remains a distinct phenomenon. While Wilson argues that this distinction may be operationally prudent [18], it can be conceptually problematic. The consensual weight of repeated assertions of the *operational* primacy of accessibility implicitly positions digital preservation *conceptually* as an essentially managerial activity, whose imperatives stop with provisioning access [19]. However, the ability to retrieve a well-managed object is distinct from a subsequent capacity to make productive use of it. The parameters of that usage are concerned with post-managerial experience.

The embrace of that experiential component recasts digital preservation as an essentially *communicative*, rather than merely managerial, enterprise. That is, it aims to facilitate future purposive human engagement with past informative expression. While that facilitation necessarily involves technological intermediation through artifactual vehicles and managerial processes, its underlying goals are fundamentally humanistic in nature. These give preeminence to the role of the information consumer [20] and the communicative outcomes of the consumer/content engagement.

The success of an act of preservation-enabled communication is dependent on its consumer-facing consequence. That is, preservation acts are successful if they result in a pertinent change to the consumer's intellectual, psychological, or physical state that otherwise would not have been known, felt, or performed [21]. As any such success is contingent with respect to time, place, person, and purpose [22], digital preservation inherently operates in an subjective sphere. Efforts to ensure beneficial outcomes over time are complicated by the fact that the passage of time is inexorably accompanied by ever-growing *technical* distance. However, the more significant preservation challenge over archival timespans is the accompanying *cultural* distance separating the points of content creation, acquisition, and use.

A communicative perspective of the digital preservation domain makes it susceptible to a *communicological* approach. Communicology is the study of individually-embodied human discourse [23], in distinction to disembodied machine-to-machine information-theoretic communication [24] and socially-embodied mass communication [25]. That discourse is viewed as a system of expressive *signs* whose meaning emerges through contingent interpretation by their consumers individually, institutionally, and culturally-positioned in socio-technical space [26]. A "sign" is a high-level abstraction for any information-laden entity that "stands to somebody for something in some respect or capacity" [27]. Stamper extended the traditional tripartite structure of a sign – syntactic form, semantic content, pragmatic experience [28] – to encompass six aspects pertinent for greater applicability to digital information systems [29] (see Table 1). Abrams proposed a seventh, performic, aspect for pertinence to digital preservation [30]. This recognizes that digital objects must be dynamically and contextually *performed* to be susceptible to analog human perception and cognitive interpretation [31][32].

The common metaphor of a digital *carrier* is the ontic (or tangibly-reified) manifestation of an abstract information-laden *message*. That message encompasses three distinct semiotic aspects:

1. Empiric symbolic encoding
2. Syntactic rhetorical expression
3. Semantic meaning or psychological affect

These generally align with the FRBR Manifestation, Expression, and Work constructs [33], which constitute an essential progression from the

Table 1. Philosophical Foundations of Digital Preservation

Concern	Managerial				Communicative		
Referent	Information object				Information experience		
Focus	Artifactual				Experiential		
Abstraction	Carrier	Message			Performance	Environment	Mind
Function	Reificatory	Representational	Rhetorical	Ontological	Epistemological	Associational	Phenomenological
Affordance	Manifestation	Encoding	Expression	Meaning	Behavior	Context	Understanding
Semiotic	Ontics	Empirics	Syntactics	Semantics	Performics	Plaistics	Pragmatics
Imperative	Integrity	Validity	Authenticity	Reliability	Accessibility	Relevancy	Legitimacy
Descriptiveness	Is-ness		Of-ness		About-ness		
Role	Enabling means				Enabled ends		
Measure	Output				Outcome		
Metric	Trustworthiness				Success		
Evaluation	Objective				Subjective		

(relatively) concrete to the (relatively) abstract. The full set of seven semiotic dimensions similarly represents a continuum of perspectives on the preservation enterprise from the objective to the subjective, spanning three descriptive categories:

1. Characteristic *is-ness*
2. Denotative *of-ness*
3. Connotative *about-ness*

This terminology is borrowed from subject cataloging theory [34], but is deployed to indicate the range of afforded descriptive scope. For example, while this paper *is* an Office Open XML document, it also is overtly descriptive *of* the derivation of a conceptual domain model for infrastructure refresh, while also being interpretatively *about* the model's novelty and legitimacy as a complement to prior modeling efforts.

Preservation outputs and outcomes are evaluated in terms of associated imperative qualities. An output is a quantifiably-measurable result of an activity, such as counts or enumerations of the generated states or productions of a system or process [35], while an outcome is a qualitatively-assessable benefit of an output [36]. That is, an outcome focuses on the experiential impact or difference an output has on the part of its recipient [37].

An ontic manifestation is *integral* if it is complete and uncorrupted [38]; an empiric encoding is *valid* if it conforms to an authoritative definition [39]; a syntactic expression is *authentic* if it expresses what it purports to express [38]; a semantic meaning is *reliable* if its factual presentation is accurate [38]; a performic behavior is *accessible* if it can be availed upon at a time and place and in a manner of the consumer's choice [40]; a plaistic context is *relevant* if it is fit for a consumer's intentional or serendipitous purpose [41]; and a pragmatic understanding is *legitimate* if it is meaningful for that purpose [42][43]. Since any given encounter with preserved digital material is dependent on time, place, person, and purpose, the consuming participant in that encounter will come to it with a potentially unique set of implicit or explicit weighting factors regarding the relative importance of these various qualities. Thus, digital preservation success should be viewed as a multi-valent evaluable factor [30].

3. Conceptual Foundations

Content analysis of digital preservation policy determines that the success of long-term digital preservation activity is commonly evaluated in terms of four normative qualities: the integrity, authenticity, accessibility, and usability of managed digital content [19]. Since these policies establish the implicit social "contract" underlying the interaction between preservation stakeholders and delegated service-providers, whether internal or external to an institutional program [30], these qualities suggest three defining imperatives for the preservation enterprise:

1. Ensuring the existence of *authentic* information *objects*
2. Supporting modalities of *authoritative* information *access*
3. Affording opportunities for *legitimate* information *experiences*

Authenticity is the quality of an object being what it purports to be; authoritativeness, that of being appropriate and reliable for the purpose at hand; and legitimacy, that of being meaningful for that contextually-situated purpose. (Authenticity is viewed as subsuming integrity, as any explicit loss of integrity inherently implies corresponding loss of authenticity.) These correspond to intentions and expectations that future preservation outcomes encompass the preserved artifact *itself*; the means to *interact with* and *know about* the artifact; and the experiential *results* of that interaction. The authenticity/legitimacy distinction contrasts objective universality (authentic for all) with subjective contingency (legitimate for one). In other words, while a given digital object is singularly either authentic or inauthentic, that same object may be susceptible to any number of legitimate (re)uses, each particular to contingent context.

Efforts to ensure these beneficial outcomes over time is complicated by the ever-increasing number, size, complexity, and diversity of digital content available for preservation attention, as well as the continual – and often disruptive – evolution and transformation of the modalities of desired (re)use. These problematic aspects of long-term stewardship can be ameliorated through a comprehensive programmatic approach to fundamental preservation concerns [44], which encompass various functional categories:

1. *Predilect* – Decide what you intend

2. *Select* – Appraise what is available
3. *Collect* – Obtain what you select
4. *Introspect* – Know what you obtain
5. *Perfect* – Enrich what you know
6. *Protect* – Steward what you have
7. *Direct* – Control how you steward
8. *Project* – Offer what you control
9. *Connect* – Provide what you offer
10. *Reflect* – Assess what you do

These extend the set of categories previously derived by Abrams [45] to provide explicit consideration of curatorial discretion regarding acquisition (selection) [46]; opportunities to augment the representation [47], description, and understanding [48] of objects, behaviors, and contexts (perfection); and programmatic governance and accountability (direction) [49]. Regarding the previously identified categories, predilection encompasses stakeholder consultation, analysis, and prioritization. Collection remains the most decisive preservation imperative: while proactive stewardship doesn't guarantee success, an absence of that stewardship almost surely guarantees failure. Introspection provides intellectual as well as technical characterization, facilitating targeted workflow development and automation. Protection lies at the artifactual core of the preservation endeavor while projection and connection mediate the experiential. Reflection supports continuous programmatic improvement.

The perspectival shift in digital preservation emphasis towards communicative information experiences suggests the desirability of similarly recasting the domain concept of significant properties to that of significant *affordances* [30]. In the preservation context, an affordance is a functional capability available to a human consumer *to do* something meaningful *with* a preserved object [50]. For example, the *property* of (quantitative) fixity *affords* the ability to determine (qualitative) integrity. Similarly, the property of an image's defined colorspace affords the ability for colorimetrically-reliable visual presentation. In other words, an affordantial perspective complements a focus on the managerial and artifactual aspects of preservation attention with communicative and experiential considerations. The experiential connotation of affordance also highlights the view of human engagement with a preserved digital object as a subjective performance [51]. The meaningfulness of the pragmatic response to such a performance is dependent on various frames-of-reference that contextualize the encounter [52]. These include the contexts of [45]:

1. Cultural production, indicative of originating creative intention
2. Curatorial appraisal, selection, and aggregation in thematic collections, through which the individual member objects accumulate associational meaning [53]
3. Prior consumption, indicative of alternative interpretive reception and response
4. Collateral lived-experience and proximate purpose of the contemporary consumer, which establish experiential expectations

While the domain concept of representation information is defined in generic terms [14], in practice it has not encompassed the means to represent, capture, and retain all of these diverse contextual positions [53]. New infrastructural systems should provide explicit support for persistent management of and experiential access to authoritative performative behaviors and relevant contextual reference frames.

4. Emergent Infrastructural Principles

Digital preservation is a complex of people, policies, procedures, as well as systems facilitating technically-mediated, but fundamentally human communication across time [54]. Given that technical infrastructure is inherently ephemeral and needs to be refreshed and re-envisioned periodically [55][56], it is appropriate to assert expansive aspirations for its function and operation during its design phase. While these may not be immediately provisionable, they set a benchmark for incrementally-achievable programmatic goals. For the Futures project, these goals include support for:

1. Any content genre, language, structure, form, number, size, and description
2. Any managerial duration (interim, persistent, or permanent) and eventuality (proactive when possible, reactive when necessary)
3. Any stakeholder competency, purpose, and modality

The first group is concerned with maximizing the scope of preservation eligibility; the second, the range of preservation intentions and expectations; and the last, the parameters of experiential (re)use.

A claim of effective support for these various goals does not necessarily imply a uniform level of outcome. Instead, effectiveness should be viewed as the condition of doing the best one can regarding a given body of digital content at a particular point in time and state of expertise, tooling, and capacity as well as controlling curatorial priority.

The design, implementation, and operation of preservation infrastructure should embrace a number of programmatically-significant qualities:

1. *Transparency* – Open decision-making [57]
2. *Stability* – Available at a time and place and in a manner of user choice
3. *Reliability* – Predictable behavior conforming to documented function [58]
4. *Productivity* – Maximal purposive impact with minimal effort
5. *Affordability* – Maximal service function at minimal total cost [59]
6. *Sustainability* – Longevity with minimal demands on necessary resources [60]
7. *Functionality* – Responsive enhancement for ever-evolving needs

These factors address important social concerns of stakeholder adoption, retention, and accountability. At a technical level, they should be complemented with other architectural principles, including:

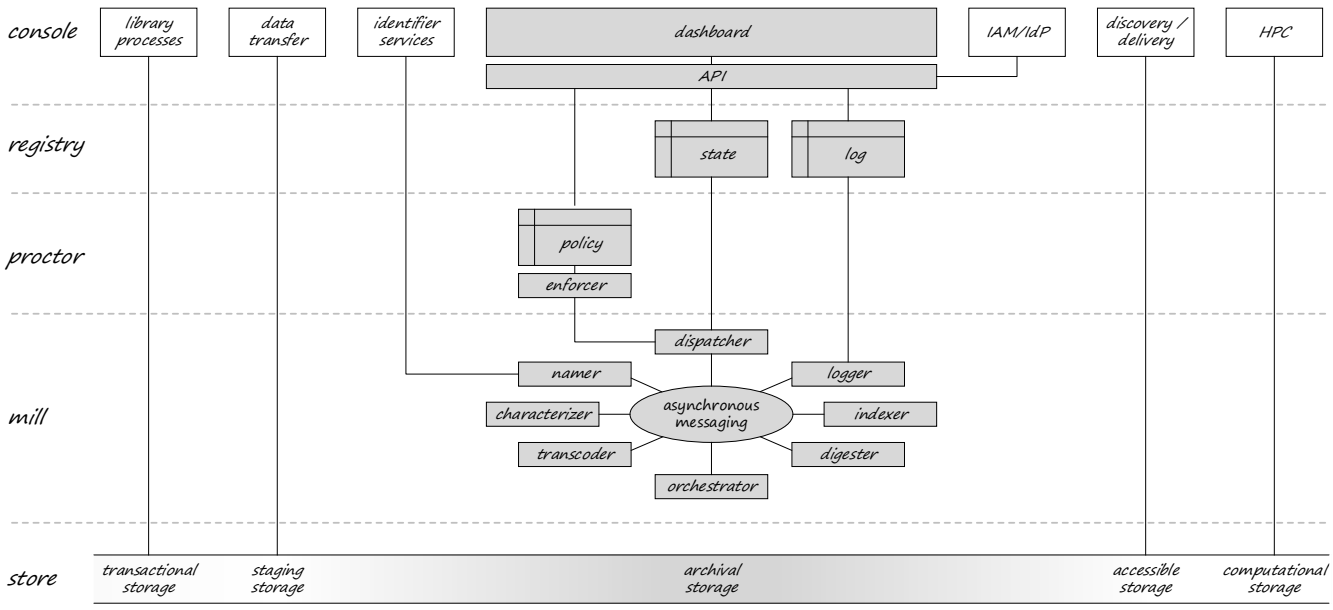
1. Separation of concerns [61]
2. Elastic scalability [62]
3. Asynchronous operation [63]
4. API-first [64]
5. Extension through (re)configuration rather than coding

The first two principles suggest an approach of decoupled interoperability through stateless microservices. The third promotes fault tolerance and adaptive error recovery with eventual consistency. The fourth ensures uniformity of function for both human and automated agents, maximizing opportunities for access modality, automation, and ecosystem integration. The final principle facilitates infrastructural sustainability and relevance through functional customization and enhancement without recourse to expensive software updates. This also permits a wider range of institutional roles to participate meaningfully in functional improvement.

Taken together, these socio-technical principles contribute to the Futures project's evolving abstract functional reference model (see Figure 1). This encompasses computational components at five tiers of abstraction:

1. *Console* – Interfaces for human and automated agents
2. *Registry* – Persistent state for content and logging of infrastructural processes
3. *Proctor* – Machine-actionable policies and automated enforcement

Figure 1. Functional reference model



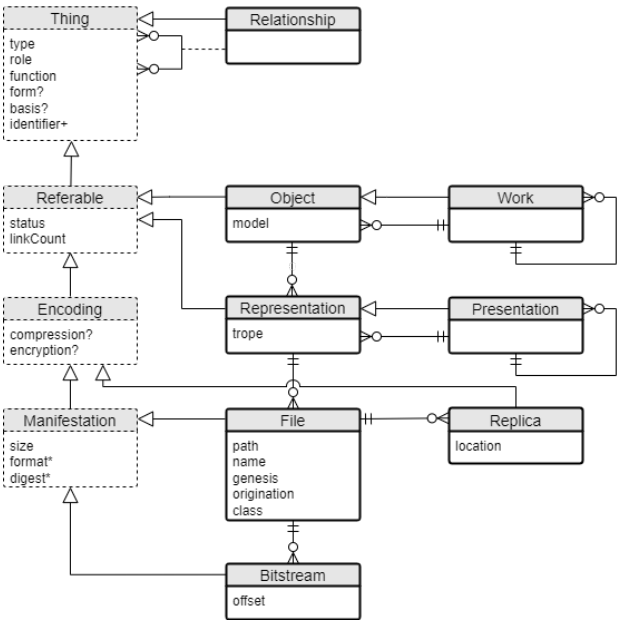
4. *Mill* – Microservice-based processing farm
5. *Store* – Bit-level persistence of tangible manifestations of content (defined by prior Library standardization on the S3 API and OCFL structuring principles [3]).

(The lower two tiers are named in playful homage to Babbage and Lovelace [65].) Note, again, that this is an *abstract* description of core functional entities and relationships. Pointedly, it is *not* intended directly as an architectural diagram or technical specification.

The core of the model conceives of ideal digital preservation infrastructure as a finite state machine. Stateful transitions are initiated by either external or internal stimuli, that is, user-specified requests such as new deposit submissions, or self-identified conditions such as fixity violations. An automated policy enforcer evaluates the stimulus in light of current content state and applicable policy rules. If necessary, the enforcer dispatches a series of potentially chained microservice invocation requests intended to bring the state back into conformance with policy prescriptions. IRODS provides a useful exemplar in this regard [66][67]. The Preservation Action Registries (PAR) initiative [68] suggests an alternative avenue of exploration regarding the expression and evaluation of policy rules.

Subsequent project activity will supplement these efforts with a stateful information model pertinent to expression and persistence of the artifactual and experiential functions, affordances, and imperatives enumerated in Table 1. The model is still under development, but its current draft form is shown in Figure 2. Its core is a four-level data hierarchy of Objects/Works, Representations/ Presentations, Files, and Bitstreams. Works define complex Object aggregations or hierarchies. Objects conform to structural/semantic content *models*, analogous to file-level MIME format typing [69]. This facilitates descriptive high-level characterization and validation, as well as aggregation of like-with-like for efficient bulk processing. Representations (defining subsets of files in the PREMIS sense [70]) and Presentations respectively model static relational file structure and dynamic navigational behavior, similar to the physical/logical distinction of a METS <fileSec> and <structMap> [71]. Similar to Object-level models, Representations are typed by characteristic *tropes* indicating their organizational structure. Files document content independent of specific instantiations, which are modeled by Replicas, similar to the FRBR Manifestation/Item distinction [33]. The Bitstream entity is introduced primarily to model the heterogeneous contents of container files.

Figure 2. Information reference model



A parallel hierarchy of abstract entities establishes common heritable properties. All are instances of the Thing ur-entity, characterized by their essential type, purposive role, informative function, and expressive form. For example, Objects are of simple or multipart type; Representations, tangible or digital type; and Files, unitary, wrapper, or container type. Similarly, Objects play a (primary) content or (operational) system role; Representations, a substantive, descriptive, or instrumental role ; and Files, a data or metadata role. Thing is subtyped to define Referable things and their status – active, (logically) deleted, (physically) purged) – and link count. The latter supports entity composition by reference as well as value. A referable Encoding documents optional compression and encryption as applied to encoded Manifestations representing formatted byte sequences.

5. Next Steps

Once the abstract reference models are fully populated, the generalized use cases and user stories synthesized from the details provided by stakeholder engagement will be aligned with the

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

DOI: 10.1145/nnnnnnn.nnnnnnn

derived cases and stories implied by the models. The consolidated cases and stories will inform the development of comprehensive functional and non-functional system and service requirements. These, in turn, will form the basis for a Request-for-Proposal (RFP) to identify plausible candidate solutions. Target candidates will be solicited from commercial vendors and community-supported open-source projects. The RFP also will be evaluated for potential internal Library software development, focusing on the integrative "gluing" together of externally-provisioned components; supplying otherwise unavailable but vital added-value function; or other areas in which the targeted allocation of institutional resources can provide a unique contribution.

6. CONCLUSION

The foundational conceptualization of a domain establishes the metaphoric as well as pragmatic boundaries of legitimate domain focus and action [72]. Current perspectives of the digital preservation enterprise promote a view largely limiting its concerns to the managerial and artifactual. While these are necessary enabling factors, they do not address sufficient attention to the communicative and experiential aspects of preservation concern. Fuller understanding and exploitation of the domain follows from complementary attention to both the enabl*ing* means as well as the enabl*ed* ends of the enterprise. The latter can be summarized as facilitating system-mediated, but fundamentally human communication unfolding across archival timespans and accompanying technical and cultural distance.

Progress towards this goal revolves around three primary digital preservation imperatives: ensuring persistence of authentic information objects; providing authoritative information access modalities; and affording opportunities for legitimate information experiences. Considerations pertinent to the first are well-examined and modeled by the broader preservation community at the abstract [14], architectural [73], and deployment [74] levels. Similar efforts regarding the second imperative are emerging through research and practice in software preservation and emulation [75]. Intentions and practices supporting the third, experiential imperative are less mature. The communicological framework proposed here provides useful structuring principles for further investigation of this final preeminent concern.

The Harvard Library DRS Futures project used this communicological framework as the basis for an open-ended exploration of the constituent components of an ideal digital preservation infrastructure. This process derived novel abstract functional and informational reference models from a set of initial axiomatic principles. While the contours of the model infrastructure are unlikely to be fully provisioned in the near term, they nevertheless constitute a critical roadmap for long-term planning of the Library's digital preservation intentions. A future phase of the Futures project will derive a constrained version of the idealized vision that is achievable and ultimately procurable and deployable. In almost all human endeavor, it is very unlikely that achievement ever exceeds aspiration. Thus, there is no reason not to set high aspirations as a benchmark for a desirable goal that can be approached incrementally. The Library hopes that its new conceptual foundation for digital preservation contributes to the success of its internal stewardship priorities, as well as provoking useful community discussion regarding the field's theoretical basis and progress towards state-of-the-art innovation and adoption.

7. REFERENCES

[1] D. Flecker. "Harvard's Library Digital Initiative: Building a first generation digital library infrastructure." *D-Lib Mag.*, vol. 6, no. 11, 2000. https://www.dlib.org/dlib/november00/flecker/11flecker.html

[2] A. Goethals, and T. Patterson. "The big migration: Lessons learned at the completion of the 10-year DRS2 project," presented at the 15th Int. Conf. on Digital Preservation, Boston, 2018.

[3] A. Woods. "A new storage paradigm for sustainable digital stewardship," presented at the CNI Fall 2022 Membership Meeting, Washington, 2022.

[4] Harvard Library and Library Technology Services. *DRS Futures Project*. https://sites.harvard.edu/drs-futures/

[5] R. Kneuper. "Sixty years of software development life cycle models," *IEEE Annals in the Hist. of Comput.*, vol. 39, no. 3, pp. 41-54, 2017, doi:10.1109/MAHC.2017.3481346.

[6] I. Jebreen. "Using inductive approach as research strategy for requirements engineering," *Int. J. of Comp. and Info. Tech.*, vol 1, no. 2, pp. 162-173, 2012. https://www.ijcit.com/archives/volume1/issue2/Paper010222.pdf

[7] K. E. Wiegers, and J. Beaty. *Software Requirements*, 3d ed., Redmond: Microsoft, 2013.

[8] A. van Lamsweerde. "Requirements engineering: From craft to discipline," presented at the 16th ACM SIGSOFT Int. Symp. on Foundations of Softw. Eng., pp. 238-259, Nov. 9-14, 2008, doi:10.1145/1453101.1453133.

[9] J. Andow. "Abduction by philosophers: Reorienting philosophical methodology," *Metaphilosophy*, vol. 47, no. 3, pp. 353-370, 2016, doi:10.1111/meta.12191.
[10] C. M. MacKenzie, J. Laskey, F. McCabe, P. F. Brown, and R. Metz. *Reference Model for Service Oriented Architecture 1.0*, OASIS, 2006. http://www.oasis-open.org/committees/tc_home.php
[11] N. C. Burbules, and B. R. Warnick. "Philosophical inquiry," in *Handbook of Complementary Methods in Education Research*, Routledge, 2006, pp. 489-502, doi:10.4324/9780203874769.ch29.
[12] K. Thibodeau. "Digital preservation," in *Encyclopedia of Archival Sci.*, L. Duranti and P. Franks, Eds., Rowan & Littlefield, 2017.
[13] *Definitions of Digital Preservation*, American Library Association, Chicago, 2009. http://www.ala.org/alcts/resources/preserv/defdigpres0408
[14] *Space data and information transfer systems – Open archival information system (OAIS) – Reference model*, ISO 14721, 2012.
[15] N. Beagrie, Ed. *Digital Preservation Handbook*, 2nd ed., Digital Preservation Coalition, Glasgow, 2015. https://www.dpconline.org/handbook
[16] *Concept of Digital Preservation*, UNESCO, Paris, 2019. https://en.unesco.org/themes/information-preservation/digital-heritage/concept-digital-preservation
[17] *Dictionary of Archives Terminology*, Society of American Archivists, Chicago, 2020. https://dictionary.archivists.org/
[18] T. C. Wilson. "Rethinking digital preservation: Definitions, models, and requirements," *Digital Library Perspectives*, vol. 33, no. 2, pp. 128-136, 2017, doi: https://doi.org/10.1108/DLP-08-2016-0029.
[19] S. Abrams. "Tacit attitudinal principles for evaluating digital preservation success," *Archival Sci.*, vol. 21, no. 3, pp. 295-315, 2021, doi: 10.1007/s10502-021-09360-5.
[20] S. Sacchi. "What do we mean by 'preserving digital information'? Towards sound conceptual foundations for digital stewardship," Ph.D. dissertation, Univ. of Illinois, 2017. http://hdl.handle.net/2142/78440
[21] R. Savolainen. "Elaborating the sensory and cognitive-affective aspects of information experience," *J. of Librarianship and Information Sci.*, vol. 52, no. 3, pp. 671–684, 2019, doi:10.1177/0961000619871595.
[22] B. W. Bishop and C. Hank. "Measuring FAIR principles to inform fitness for use," *Int. J. of Digital Curation*, vol. 13, no. 1, pp. 35-46, 2018, doi:10.2218/ijdc.v13i1.630.
[23] R. L. Lanigan. "Philosophy of communicology: 'Discourse which embodies itself is communication'," *Review of Communication*, vol. 15, no. 4, pp. 349-358, 2015, doi:10.1080/15358593.2015.1102408.
[24] R. L. Lanigan. "Communicology," in *Int. Encyclopedia of Communication*, W, Donsbach, Ed., Blackwell, 2013, doi:10.1111/b.9781405131995.2008.x.
[25] I. E. Catt. "The two sciences of communication in philosophical context," *Review of Communication*, vol. 14, no. 3-4, pp. 201-228, 2014, doi:10.1080/15358593.2014.986876.
[26] J. Mingers and L. WIllcocks. "An integrative semiotic methodology for IS research," *Information and Organization*, vol. 27, no. 1, pp. 7-36, 2017, doi:10.1016/j.infoandorg.2016.12.001.
[27] C. S. Peirce. *Collected Papers of Charles Sanders Peirce. Volumes I and II: Principles of Philosophy and Elements of Logic*, Cambridge: Harvard University Press.
[28] C. Morris. *Signification and Significance*, Cambridge: MIT Press, 1964.
[29] R. K. Stamper. "A semiotic theory of information and information systems," presented at the ICL/University of Newcastle Seminar on 'Information', 1993 https://assets.cs.ncl.ac.uk/seminars/101.pdf
[30] S. Abrams. "A communicological critique of evaluative norms for digital preservation success," Ph.D. dissertation, Queensland Univ. of Tech., Brisbane, 2023, doi:10.5204/thesis.eprints.238194.
[31] H. Heslop, S. Davis, and A. Wilson. *An Approach to the Preservation of Digital Records*, National Archives of Australia, 2002. https://www.naa.gov.au/sites/default/files/2020-01/An-Approach-to-the-Preservation-of-Digital-Records.pdf
[32] C. Becker. "Metaphors we work by: Reframing digital objects, significant properties, and the design of digital preservation systems," *Archivaria*, vol. 85, 2018. https://archivaria.ca/index.php/archivaria/article/view/13628
[33] P. Riva, P. Le Bœuf, and M. Žumer. *FRBR-Library Reference Model*, IFLA, 2016. https://www.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr-lrm/frbr-lrm_20160225.pdf
[34] B. Hjørland. "Subject (of documents)," *Knowledge Organization*, vol. 44, no. 1, pp. 55-64, 2017. http://www.isko.org/cyclo/subject
[35] R. E. Dugan and P. Hernon. "Outcomes assessment: Not synonymous with inputs and outputs," *J. of Academic Librarianship*, vol. 28, no. 6, pp. 376-380, 2022, doi:10.1016/S0099-1333(02)00339-7.
[36] M. Kyrillidou. "From input and output measures to quality and outcome measures, or, from the user in the life of the library to the library in the life of the user," *J. of Academic Librarianship*, vol. 28, no. 1-2, pp. 42-46, 2002, doi:10.1016/S0099-1333(01)00299-3.
[37] G. Tsakonas and C. Papatheodorou. "An ontological representation of the digital library evaluation domain," *J. of the American Society for Information Sci.*, vol. 62, no. 8, pp. 1577-1593, 2011, doi:10.1002/asi.21559.
[38] J. Kastenhofer. "The logic of archival authenticity: ISO 15489 and the varieties of forgeries in archives," *Archives and Manuscripts*, vol. 43, no. 3: pp. 166-180, 2015, doi: 10.1080/01576895.2015.1074085.
[39] M. Lindlar and Y. Tunnat, "How valid is your validation? A closer look behind the curation of JHOVE," *Int. J. of Digital Curation*, vol. 12, no. 2, pp. 286-298, 2017, doi: 10.2218/ijdc.v12i2.578.
[40] L. Jaillant. "How can we make born-digital and digitized archives more accessible?" *Archival Sci.*, vol. 22, no. 3, pp. 417-436, 2022, doi:10.1007/s10502-022-09390-7.
[41] C. A. Lee. "A framework for contextual information in digital collections," *J. of Documentation*, vol. 67, no. 1, pp. 95-143, 2011, doi:10.1108/00220411111105470.
[42] C. Dallas. "An agency-oriented approach to digital curation theory and practice," presented at the International Cultural Heritage Informatics Meeting, Toronto, 2007. http://www.archimuse.com/ichim07/papers/dallas/dallas.html
[43] L. Duranti and K. Thibodeau. "The concept of record in

interactive, experiential and dynamic environments: The view of InterPARES," *Archival Sci*., vol. 6, no. 1, pp. 13-68, 2006, doi:10.1007/s10502-006-9021-7.
[44] E. Baucom. "Planning and implementing a sustainable digital preservation program," *Library Technology Reports*, vol. 55, no. 6, pp. 5-28, 2019, doi:10.5860/ltr.55n6.
[45] S. Abrams. "A foundational framework for digital curation: The Sept domain model," presented at the 12th Int. Conf. on Digital Preservation, Chapel Hill, 2015. https://hdl.handle.net/11353/10.429533
[46] L. Work, N. Tallman, M. Shallcross, and C. Mumma. "The right stuff (over time)," presented at the NDSA Digital Preservation Conf., 2020, doi:10.17605/OSF.IO/875XG.
[47] T. Owens. *The Theory and Craft of Digital Preservation*, Baltimore: Johns Hopkins Univ. Press, 2018.
[48] T. Hutchinson. "Natural language processing and toolsets for archival processing," *Records Manage. J.*, vol. 30, no. 2, 2020, doi:10.1108/RMJ-09-2019-0055.
[49] C. Becker, G. Antunes, J. Barateiro, and R. Vieira. "A capability model for digital preservation," presented at the 8th Int. Conf. on Preservation of Digital Objects, Singapore. https://www.ifs.tuwien.ac.at/~becker/pubs/becker_ipres2011.pdf
[50] M. Hedstrom and C. A. Lee. "Significant properties of digital objects: Definitions, applications, implications," presented at the DLM Forum, Barcelona, 2002, pp. 218-227. https://ils.unc.edu/callee/sigprops_dlm2002.pdf
[51] A. Dappert and A. Farquhar. "Significance is in the eye of the beholder," presented at the 13th Eur. Conf. on Res. and Adv. Tech. for Digital Libraries, Corfu, 2009. http://www.planets-project.eu/docs/papers/Dappert_SignificantCharacteristics_ECDL2009.pdf
[52] M. Bonn, L. Kendall, and J. McDonough. "Preserving intangible heritage: Defining a research agenda," *Proc. Of the Assoc. for Information Sci. and Tech.*, vol. 53, no. 1, 2016, pp. 1-5, doi:10.1002/pra2.2016.14505301009.
[53] H. Brocks, A. Kranstedt, G. Jäschke, and M. Hemmje. "Modeling context for digital preservation," in *Smart Information and Knowledge Management*, e. Szczerbicki and N. T. Ngugen, Eds., Springer, 2010, pp. 197-226, doi:10.1007/978-3-642-04584-4.
[54] S. Abrams. "Theorizing success: A communicological approach to evaluating digital preservation efficacy," *Bull. of IEEE Technical Committee on Digital Libraries*, vol. 15, no. 1, 2019. https://bulletin.jcdl.org/Bulletin/v15n1/papers/abrams.pdf
[55] G. Janée, J. Frew, and T. Moore. "Relay-supporting archives: Requirements and progress," *Int. J. of Digital Curation*, vol. 4, no. 1, 2009, doi:10.2218/ijdc.v4i1.78.
[56] J. Barateiro, G. Antunes, F. Freitas, and J. Borbinha. "Designing digital preservation solutions," *Int. J. of Digital Curation*, vol. 5, no. 1, 2010, doi:10.2218/ijdc.v5i1.140.
[57] D. Lin et al. "The TRUST principles for digital repositories," *Sci. Data*, vol. 7, no. 144, 2020, doi:10.1038/s41597-020-0486-7.
[58] E. M. Corrado. "Repositories, trust, and the CoreTrustSeal," *Technical Services Quarterly*, vol. 36, no. 1, pp. 61-72, 2019, doi:10.1080/07317131.2018.1532055.
[59] N. Tallman and H. Wang. "Seeking sustainability," presented at the 18th Int. Conf. on Digital Preservation, Glasgow, 2022, doi: 10.26207/58rw-kt80.
[60] K. Pendergrass, W. Sampson, T. Walsh, and L. Alagna. "Towards environmentally sustainable digital preservation," *American Archivist*, vol. 82, no. 1, 2019, pp. 165-206: doi:10.17723/0360-9081-82.1.165.
[61] H. Ossher and P. Tarr, "Multi-dimensional separation of concerns and the hyperspace approach," in *Softw. Arch. And Component Tech.*, M. Aksit, Ed., Springer, 2002, pp. 293-324.
[62] D. M. Gerrard, J. E. Mooney, and D. Thompson. "Digital preservation at Big Data scales," *Library Hi Tech*, vol. 36, no. 3, pp. 524-538, 2018, doi:10.1108/LHT-06-2017-0122.
[63] R. Laigner, Y. Zhou, M. A. V. Salles, and Y. Liu. "Data management in microservices," *Proc. Of the VLDB Endowment*, vol. 14, no. 13, pp. 3348-3361, 2021, doi:10.14778/3484224.3484232.
[64] N. Beaulieu, S. M. Dascalu, and E. Hand. "API-first design: A survey of the state of academia and industry," presented at the 19th Int. Conf. on Information Tech. – New Generations, 2022, pp. 73-79, doi:10.1007/978-3-030-97652-1_10.
[65] W. Dickey. "The mill and the store," *New England Review and Bread Loaf Quarterly*, vol. 10, no. 1, pp. 98-111, 1987. https://www.jstor.org/stable/40241889
[66] M. Conway, R. Moore, A. Rajasekar, and J.-Y. Nief. "Demonstration of policy-guided data preservation using IRODS," presented at the IEEE Int. Symp. On Policies for Distributed Systems and Networks, Pisa, 2011, pp. 173-174, doi:10.1109/POLICY.2011.17
[67] M. C. Conway. "Policy domains: A state-machine based approach to policy-based data management," M.S. thesis, Univ. of North Carolina, Chapel Hill, 2017, doi:10.17615/3330-ke07
[68] J. O'Sullivan and J. Tilbury. "Using preservation action registries to automate digital preservation," *J. of Digital Media Manage.*, vol. 9, no. 3, pp. 240-252, 2021. https://www.ingentaconnect.com/content/hsp/jdmm/2021/00000009/00000003/art00007
[69] A. Kirchhoff and S. Morrissey. "Digital preservation metadata practice for e-journals and e-books," in *Digital Preservation Metadata for Practitioners*, A. Dappert, R. S. Guenther, and S. Peyrard, Eds., Springer, 2016, pp. 83-97, doi:10.1007/978-3-319-43763-7_7
[70] P. Caplan. *Understanding PREMIS*, Library of Congress, Washington, 2009; rev. 2021. https://www.loc.gov/standards/premis/understandingPREMIS_english_2021.pdf
[71] *Metadata Encoding & Transmission Standard*. Library of Congress, Washington, 2010. https://www.loc.gov/standards/mets/METSPrimer.pdf
[72] P. B. Condon. "Digital curation through the lens of disciplinarity in the development of an emerging field," Ph.D. dissertation, Simmons Univ., Boston, 2014. https://beatleyweb.simmons.edu/scholar/files/original/c79af2ebd58dfb5cdaf3a26d95f03419.pdf
[73] N. Tallman. "A 21st century technical infrastructure for digital preservation," *Information Tech. and Libraries*, vol. 40, no. 4, 2021, doi:10.6017/ital.v40i4.13355.
[74] *Core Requirements for a Digital Preservation System*, Digital Preservation Coalition, Glasgow, 2022. https://www.dpconline.org/docs/knowledge-base/2581-core-requirements-for-a-digital-preservation-system-v1
[75] E. Cochrane, K. Rechert. J. Oberhauser, S. Anderson, C. Fox, and E. Gates. "Useable software forever," presented at the 18th Int. Conf. on Digital Preservation, Glasgow,

2022, pp. 40-52. https://www.dpconline.org/docs/miscellaneous/events/2022-events/2791-ipres-2022-proceedings/file

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

AROUND FOR DECADES, GONE IN A FLASH

How we dealt with Flash objects at the National Archives of the Netherlands

Remco van Veenendaal
National Archives of the Netherlands
Netherlands
remco.van.veenendaal@nationaalarchief.nl
0000-0002-2351-1677

Lotte Wijsman
National Archives of the Netherlands
Netherlands
lotte.wijsman@nationaalarchief.nl

Jacob Takema
National Archives of the Netherlands
Netherlands
jacob.takema@nationaalarchief.nl

Marin Rappard
National Archives of the Netherlands
Netherlands
marin.rappard@nationaalarchief.nl

Abstract – In 2020, Adobe announced that they would end support for Adobe Flash Player. Initially, we (the preservation team at the National Archives of the Netherlands) assumed we had only a few or no Flash objects in our digital repository, but this assumption turned out to be incorrect. The discovery of Flash objects in our holding led to the start of a research project to answer several questions. Through a series of dedicated meetings, we formulated a strategy focused on preserving ongoing accessibility to our Flash objects through emulation. We were curious to find out *if* we had Flash objects, *where* they were located, and *which solution* would help us render these objects. This was done with the use of the three preservation functions (Watch, Action, and Planning). After locating the Flash objects, we were able to test potential solutions. The results were then applied to our situation at the National Archives. This led to the development of conclusions and several pieces of advice accompanying those.

Keywords – Flash, emulation, migration
Conference Topics – From Theory to Practice

I. INTRODUCTION

At the National Archives of the Netherlands (NANETH), we have implemented the three important preservation functions, namely: Preservation- Watch, Planning, and Action. Team preservation NANETH uses Watch to undertake an in- and external risk assessment. These risks can be changes in the technical environment, the user community, and organization (e.g., budget cuts). Planning then allows us to develop advice for previously identified risks. This can be done in collaboration with potential stakeholders. Our advice will then be transferred to the collections department, who are responsible for its implementation at NANETH. [1]

Although rumors about Adobe Flash's impending End of Live status had been circulating for a while, it was in September 2020 that our team discovered the news that Adobe would definitively end support of Adobe Flash Player on December 31st, 2020. Adobe would also block Flash objects from running. [2] Initially, we thought the impact to our holdings to be minimal, expecting no or only a few information objects containing Flash to be present in the collection. However, a quick scan of our holdings showed that we did have Flash objects in our digital repository. This led us to change our risk assessment (Watch) of Flash from 'no risk' to 'potential risk'. Before starting the Planning function, we formulated several research questions concerning the subject of Flash content in our archives:

- How many Flash objects do we have in our digital repository?
- Is the assumption correct that the Flash content can mainly be found in web archives/websites?

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

- What is the impact of the Flash content, and can Flash be rendered in pywb[1] or another viewer?
- What are possible strategies for keeping Flash sustainably accessible?

Our goal was to identify the magnitude of the problem, the potential solutions, and selecting which one would suit our organization best. Through a series of meetings dedicated to researching Flash, we eventually formulated a strategy or advice for preserving our Flash objects in a way that ensures their ongoing accessibility through emulation. Setting up several dedicated meetings, we ensure within our team that there is opportunity to work on these extended projects with the entire team. This way, we can learn from and with each other while also working toward a final product (e.g., an advice, research report).

II. What is Adobe Flash?

Adobe Flash is a software platform that allows for animations, web videos, and web application (e.g., games and websites) to be created. It was primarily used to design websites and advertisements on websites, also known as banners. Subsequently, Adobe Flash Player is the viewer that could be used to view the content that we created with Adobe Flash. Flash had an immense user base for creating interactive websites at first. However, with the introduction of HTML5 this decreased. Moreover, security issues were identified, which led Adobe to transfer to the Adobe Air platform. Flash Player was eventually deprecated in 2017 and became end-of-life in 2020 for all users outside of China and the non-enterprise users. [3]

III. Finding Flash

We used several methods to answer the question of how many Flash objects are present in our digital repository. We conducted our initial search for Flash objects by extension. We had already found out that, while there are more options for Flash objects, the extensions .fla, .swf, and .flv were most relevant to our holdings.

- The Macromedia Flash FLA Project File Format with .fla extension is the 'authoring' format for the application software. It's a proprietary format and therefore only able to be created and edited in Adobe Animate and Adobe Flash Pro. Objects in this format contain the original, uncompressed source files for Flash animations and applications and are used to store vector graphics, pictures, text, animation timelines, and other components necessary to make a Flash project. They also include metadata such as project settings and scripting code required to provide interactivity and other project abilities.
- For distribution, the 'final result' of these FLA project files is typically exported to a Shockwave Flash file, the compiled format for sending Flash content over the internet. SWF files are formed by assembling and compressing the FLA file's assets and elements. The assembled SWF (pronounced 'Swiff') file includes all of the information required to show and interact with the content, such as the timeline and stage attributes. SWF files can include complex features such as scripting, vector graphics, and multimedia playback in addition to animations and interactive content. The SWF files are compressed and optimized to reduce their size, which results in them not being easily modified or edited.
- Alternatively, FLA projects can also be exported to Flash Video or .flv files. This is a video container that supports a variety of video codecs and several audio codecs. These files can still be opened with software such as Adobe Animate (multiplatform), Media Player Classic (Windows), VideoLAN VLC media player (multiplatform) and individual objects in this format, depending on the codecs used, might therefore be less 'at risk' than previously mentioned formats.[2]

Unfortunately, simply searching by these extensions was not foolproof. It resulted in giving us false hits in addition to giving us valid results. This was due to the fact that our digital repository, obviously, also considered text containing our search terms as a hit. By using an added filter, we were able to fill in the

[1] Pywb is a Python web archiving toolkit for replaying web archives. From: https://github.com/webrecorder/pywb.

[2] See PRONOM and https://www.loc.gov/preservation/digital/formats/fdd/fdd000132.shtml.

search terms in the File name field. This gave us exclusively Flash objects. However, we are aware that extension is not a solid guarantee for finding file formats. For this reason, we use PRONOM and Digital Record Object Identification (DROID) in our digital repository. Using the PRONOM Unique ID (PUID), we assembled a list to search NANETH's digital repository for Flash objects. This resulted in the following search query:

"x-fmt/382" OR "fmt/507" OR "fmt/757" OR "fmt/758" OR "fmt/759" OR "fmt/760" OR "fmt/671" OR "fmt/762" OR "fmt/763" OR "fmt/764" OR "fmt/765" OR "fmt/766" OR "fmt/767" OR "fmt/768" OR "fmt/769" OR "fmt/770" OR "fmt/771" OR "fmt/772" OR "fmt/773" OR "fmt/775" OR "fmt/776" OR "fmt/505" OR "fmt/506" OR "fmt/104" OR "fmt/105" OR "fmt/106" OR "fmt/107" OR "fmt/108" OR "fmt/109" OR "fmt/110"

After this advanced search, we discovered that we do have Flash objects in our digital repository, at two levels: as single objects, in a folder structure of a website, and as objects in ZIPs or WARCs, as part of a harvested website. The search yielded nine results:

- Four separate objects
- One ZIP-file
- Five WARCs

With the ZIP-file and the WARCs, we had now discovered they contained Flash objects, but not how many and that they contained. Further research outside our digital repository resulted in figuring out there were three Flash objects in the ZIP-file and a total of nine in the WARCs. In addition to not being able to directly query our digital repository to find out how many Flash objects we have, we also didn't know the exact location of the Flash objects within those containers. To figure this out, you have to look inside the containers, by unzipping them (ZIP), for example, or creating indexes for them (WARC). Therefore, we downloaded the ZIP-files to our laptops to look inside the map structure. With the WARC-files, we downloaded them to our laptops so we could open the files with Notepad++ (a source code editor). [4] Opening the WARC files in Notepad++ allowed us to look at the entire WARC and the building blocks within it. By using Ctrl + F we could search for the extensions previously identified (.fla, .swf, and .flv). This will show us where the Flash content is present within the website and gives a slight indication to what it is about.

In addition to not being able to directly query our digital repository to find out how many Flash objects we have, we also didn't know the exact location of the Flash objects within those containers. To figure this out, you have to look inside the containers, by unzipping them (ZIP), for example, or creating indexes for them (WARC).

In total, we found four individual objects, three objects in a ZIP-file, and five WARCs, bringing us to a total of nine objects. Among the Flash objects were several interactive maps. We also found audio files and headers that were loaded into an interactive Flash object. In absolute terms this may not sound like a big problem, but at the time of this search there were about ten ZIPs and 25 WARCs in the digital repository. Relatively speaking, a tenth of our ZIPs and a fifth of our archived websites were in danger of information loss. Since governmental organizations have a period of 20 years to transfer archival records not selected for destruction to NANETH, we can only expect these numbers to grow in the coming years. This idea was further strengthened after our more specific research into the Flash objects in our repository showed that one of the objects was merely a reference to another website. The web page, part of the website of the minister of the Interior and Kingdom Relations of Netherlands, shows a small article that warns against bicycle theft (a very important issue in a country with more bicycles than people). Accompanying the article is a hyperlink to a video (the Flash object in question). However, this video is not present on the harvested website, but on the website of another ministry. This ministry has not yet transferred their web archive to us, so we can expect this video in the next couple of years in our digital repository.

Our second question during this stage was if the assumption was correct that our Flash content could solely be found in web archives/websites. After our search, this assumption was found to be correct. The separate Flash objects are located in the folder structure of an archived website, while the ZIP files are a compressed website. It is still possible that future transferred archival records with Flash objects will not be limited to websites. They could for example be cd-roms with Flash animations in government campaigns, raising the public's

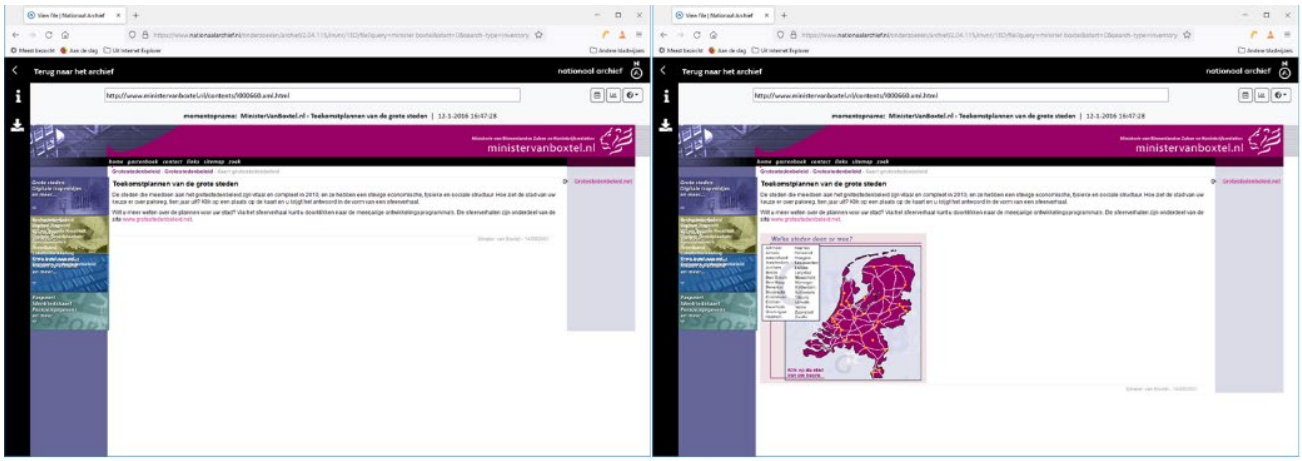


Figure 1. Without the Flash content (left) and with Flash content (right).

awareness of some topic. However, the current situation and short-term predicted situation is limited to websites.

Having identified the Flash content within our digital repository, we were able to establish the impact further by trying to view the content. A good example was an archived webpage of the website of the Minister of Metropolitan- and Integration Policy, Roger van Boxtel. The website was archived in 2016 from servers, the date on the website is August 22nd, 2002.[3] The Flash object on this page is a map of the Netherlands wherein larger cities can be selected. When selecting one of the cities, the user is then sent to a story about future plans for that particular city.[4] Fig. 1 shows the impact of not being able to view the Flash content on the left. On the right, the user is able to see the website as a whole, including Flash content. As you can see in fig. 1, we did eventually succeed in being able to see the Flash content. In chapter III we will elaborate on this more.

Users can of course use (website) viewers at home. However, these are not equipped to show Flash content by default. Moreover, our users do not have the information that we have Flash content in our web archive, how much there is, where they are situated exactly, and what the best solution is to view these objects. This gives us two possible solutions:

- NANETH-side solution: implementation on the side of NANETH. This allows our users to view our web archive without investing time to research how to view it and subsequently installing that solution.
- User-side solution: implementation by the user. We find this to be less ideal since it expects a certain degree of research and effort on the side of the user. At the National Archives we are committed to and stand for low-threshold sustainable accessibility. If you expect your users to install viewers, you create barriers. This is also why we find a client-side solution for Flash in Web sites/web archives undesirable.

IV. Dealing With Prejudice

The two preservation strategies considered for keeping Flash content accessible were file format migration and emulation. With migration you migrate the information from an older or less durable file format to a more modern or durable file format. Emulation allows you to mimic the old hardware and software environments in a modern hardware and software environment. [5] Our preservation policy doesn't explicitly state a preferred choice between the two. However, our daily practice shows a clear inclination towards migration. This is due to, for example, the technical and legal challenges included in emulation. This led

[3] https://www.nationaalarchief.nl/onderzoeken/archief/2.04.115/invnr/1ED/file?eadID=2.04.115&unitID=1ED

[4] The website the map links to, www.grotestedenbeleid.net, is no longer online. However, it has been archived by the Internet Archive. Thanks to the map's link, users can find, for example, https://web.archive.org/web/20030516013940/https://www.grotestedenbeleid.net/www/sfeermenu/sfeer/amsterdam/index.html (accessed 6-3-2023) there. Without the map, without Flash, users of the archived website of Minister van Boxtel miss the link between the cities of the map and the atmospheric stories of the metropolitan policy.

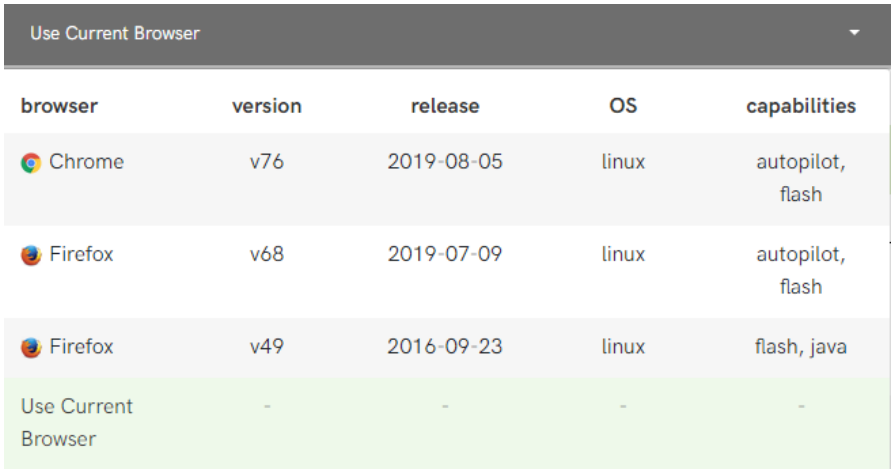


Figure 2. Conifer gives the user the option of multiple browser-versions. As seen in the figure, both Google Chrome v76 and Firefox v68 have the capability to render Flash objects.

us to start with investigating migration as a potential solution. However, to be able to properly assess potential solutions, we first needed to see the Flash content rendered to know exactly what we are

dealing with.

A. Rendering

To render the Flash content, we tested three ways:

- Conifer [6]
- Ruffle [7]
- Browsers with older version of Flash Player [8]

Previously, we mentioned our inclination toward migration. However, these three ways are all emulation-based. Our initial searches did not yield any migration-based solutions. At this time, we started to realize that our inclination towards migration would be based on the previously mentioned outdated prejudice that surrounds emulation.

In addition to using our own collection for these tests, we also used a collected corpus of Flash objects. This corpus was compiled with, among others, the use of the Internet Archive, the UK Web Archive, and the Apache Software Foundations test sets. [9]

Conifer allowed us to launch an environment containing an emulated older browser with Flash support. Opening the website of Minister van Boxtel in that browser shows the interactive map, as seen in fig. 1. The benefits of using Conifer are that it is open-source, allows the use of multiple browser-versions (see fig. 2), and is free to use. The disadvantages are that you need to register as a user, and the limit set on the amount of concurrent users, which results in waiting times. With Conifer being an online service, this solution would be a user-sided solution. Furthermore, it is unclear to us to what extent an emulated browser passes the security risks associated with Flash Player to the user's computer. However, as long as the emulated browser is offered only for trusted Flash objects from our collection, this risk will be negligible.

Using the Ruffle website, we were able to download a Flash emulator. This standalone version allows the user to render loose Flash objects, while the browser plugin shows the objects in the websites themselves. Using the plugin, we were able to render the interactive map. As with Conifer, Ruffle is open-source and free to use. Ruffle also has the benefit of

Solution	Open source	Installation required	Client- or server-sided
Conifer	Yes	No	Client
Ruffle	Yes	Yes	Client
Browsers with older version of Flash Player	No	Yes	Client

Table 1. Overview of the results of the three solutions tested.

being available as either a standalone version as a browser plugin. However, both of these need to be installed by the user. Therefore, it is a user-sided solution, which is not preferred by us.

Using the Internet Archive, you can download versions of both Mozilla Firefox and Google Chrome that have an older version of Flash Player installed. This third option, like the previous two, also provided us with a rendering of the interactive map. This solution allows the user to view the Flash objects in their 'original environment' using a free download. Nonetheless, as with Ruffle, it is a user-sided solution. Moreover, the risks associated with Flash Player are brought in.

The three solutions each have their own pros and cons connected to them. Table 1 shows an overview of a few findings connected to these solutions.

B. File Format Migration

When we investigated file format migration as a strategy, it quickly became apparent that there is little to no open-source tooling available to migrate Flash objects. NANETH's digital repository offers no tooling for it, and the leading registries in our field such as PRONOM [10], WikiData for Digital Preservation [11], and the Community Owned digital Preservation Tool Registry (COPTR) [12] also don't mention any Flash migration tools. There used to be Google Swiffy, a Flash to HTML5 converter, but this tool was more specifically intended to work on banner advertisements and has since been taken down. [13] There are other commercial providers, but these are expensive and we discovered other reasons why migration does not seem to be the best approach in NANETH's case.

Desk research led us to conclude that there are two main approaches for migrating Flash objects. Interactive Flash objects are often migrated to HTML5. Flash movies can also be migrated to MP4. [14] The Flash objects in our digital repository are primarily .swf files. These are the distributable "compiled" versions of the Flash objects, not the "source code" underlying them. Since there are no known conversion paths from SWF to HTML5, the objects would usually have to be redeveloped from scratch. As Maheswari and Reddy show in their article, the time this takes per object varies from 1 hour to 51 hours. We would assume few organizations will have the resources to allow this as a preservation strategy, depending on the scale of Flash content present. [15] However, even then, it would require specific skills to completely rebuild the look and feel of these objects. Maheshwari and Reddy argue this is because;

- "Recreation of Flash assets like images, vector graphics and animations while adhering to all the aesthetic details is a resource intensive effort.
- Developers may often lack the domain knowledge required for the particular animation. Therefore, they will have to perform the additional step of enumerating all the animation states, before rewriting the entire logic in JavaScript which in itself is a huge task." [14]

An additional complicating factor with migration is that we found many of our Flash objects within archived websites: inside a ZIP-file or WARC. Even if we did manage to migrate those objects into different formats, we would have to unpack the containers, modify all references to the migrated Flash objects, and then repackage the containers. This is a laborious process, requiring us to modify not only the Flash object, but also the container in which it is packaged.

Eventually we came to the conclusion, to our surprise, that migrating our Flash objects did not seem to be the most suitable strategy. This was affirmed during our research of tools to display WARCs that contain Flash objects, where we came across interesting alternatives.

C. *Emulation*

We found that there were ways to render Flash objects available in the open-source domain and often emulation-based. Thereby, solutions are available, such as Ruffle, which do not have the same security problems as Flash Player does. According to Ruffle, it is even possible to offer Ruffle as a server-side solution and embed Ruffle into Web pages containing Flash objects via JavaScript. [16] The user then does not have to install anything themselves. A test with the website of Minister van Boxtel on a test web server showed that this can indeed be realized with little work. However, this did require modifying the web page that contains the Flash object.

A solution that seems to fit our infrastructure even better is related to the Conifer solution previously mentioned. At NANETH we connected the Webrecorder Python wayback [17] web archiving toolkit to our digital repository for the playback of web archives on our website. In 2020 the first incrementally harvested web archives were transferred to NANETH. This led to the first (Agile) user stories calling for a web archive viewer with support for this type of web archives. After some research, we chose to implement pywb. The older browser emulation functionality that has been built into Conifer is available to install in pywb environments and is called pywb remote browsers [18]. This solution allows you to provide an emulated older browser version with Flash support. [19]

An additional advantage of being able to provide emulated older browser versions is, that it allows us to display other archived websites, without Flash objects, in a browser that was common when the website still existed. Developments in browser technology and Internet standards can cause modern browsers to display older websites differently than older browsers, whereby the rendering of the older websites in the "natural habitat" of the older browser may yield a more authentic result.

V. TAKING ACTION

As shown in the previous chapter, emulation, for us, is the preferred preservation strategy for Flash objects. Emulation solutions are readily available and available in the open-source domain. They are even available for our pywb infrastructure.

Our advice for action to be taken is threefold. The first advice is a prerequisite for the other two. Without it, the others cannot be realized.

1. We need to document and/or create metadata in which web archives Flash objects are present.[5] That will allow us at NANETH to inform our users about it, while activating server-side solutions for these archives with Flash objects.
2. Short-term advice: inform the user that they are viewing an archived website that contains Flash objects. Additionally, provide instructions on the actions they would need to take to view the content, like downloading an emulator or plug-in such as Ruffle. This is a user-sided solution, since it requires time and effort from our users.
3. Long-term advice: A server-sided emulation solution would need to be integrated into our infrastructure, so users are able to view our Flash objects without investing their own time and effort. Pywb remote browsers is an example of this server-sided emulation solution. While we can get this started in the immediate future, necessary prioritization and lead time will mean it will take longer to realize than the second advice.

These three pieces of advice have been forwarded to our collections department, together with our research report, so they can implement the most suitable solution. Team preservation is part of another department that serves as a sort of consultancy branch. This means we have preservation advisors present in our team, not acquirers or custodians of our collection. For this

[5] Currently we only have Flash content in our web archives. In time, we can of course receive other Flash content that lies outside our web archives.

reason we have forwarded our advice, so that they can implement a solution.

While our collections department can implement the first two solutions mostly on their own, with possibly some help needed for adjustments to our website, the third advice needs further work that involves multiple departments. User stories will need to be created that work into our continuous project concerning web archiving. Subsequently, our IT department needs to implement this solution.

In chapter III, we briefly mentioned the security risks associated with Flash Player. In both types of solutions (user- and NANETH-sided) this has to be considered at all stages. We as the National Archives, after all, cannot afford to have our users install an unsafe plugin, or send potentially unsafe content to the user's browser. Therefore, we have explicitly stated this in our report.

VI. CONCLUSION

Flash has been phased out and is no longer supported by default. We have only a few Flash objects in our collection at the moment, but relatively speaking, a tenth of our ZIPs and a fifth of our WARCs are at risk of information loss. We expect that our collection of web archives is going to and will continue to grow substantially. However, the phasing out of Flash is a significant risk for sustainable accessibility, especially when transferring legacy websites. In post-2020 websites, we expect to find little to no Flash.

We found that emulation is a better strategy than migration for rendering Flash objects. Our research resulted in three pieces of advice, which can be realized in stages: document the presence of Flash objects in our collection, inform the user about the presence of Flash and solutions to render Flash objects, and develop a server-side solution for displaying Flash objects.

Each realized advice reduces the risk of information loss. After the realization of our long-term advice (no. 3), web archives with Flash objects can be authentically rendered. The choice of pywb remote browsers allows other Web archives to be displayed in older browsers, which can also benefit their authentic display.

This research taught us a lot about Flash and the object present in our digital repository. This knowledge will help our organization to not only deal with the objects already present, but also with potential future Flash objects transferred to our digital repository. As mentioned previously, our team is part of the 'consultancy branch' at NANETH. We will also use our Flash research to give advice to other governmental organizations and archival institutions.

1. REFERENCES

[1] Pepijn Lucker, Remco van Veenendaal, Marcel Ras, and Barbara Sierman, "Preservation Watch at the National Archives of The Netherlands," iPRES 2018: https://doi.org/10.17605/OSF.IO/KQ6HM.

[2] https://www.adobe.com/products/flashplayer/end-of-life.html. Adobe. "Adobe Flash Player EOL General Information Page." Last modified January 13th, 2021.

[3] https://en.wikipedia.org/wiki/Adobe_Flash. Wikipedia. "Adobe Flash." Accessed February 28th, 2023; https://en.wikipedia.org/wiki/Adobe_Flash_Player. Wikipedia. "Adobe Flash Player." Accessed February 28th, 2023.

[4] https://notepad-plus-plus.org/. Notepad++. Accessed June 25th, 2023.

[5] For the definition of file format migration we used: https://www.archives.govt.nz/manage-information/how-to-manage-your-information/digital/file-format-migration. Te Rua Mahara o te Kāwanatanga Archives New Zealand. "File format migration." Accessed March 3rd, 2023; For the definition of emulation we used: https://www.dpconline.org/handbook/glossary. Digital Preservation Coalition. "Glossary." Accessed March 3rd, 2023.

[6] https://conifer.rhizome.org/. Conifer. Accessed March 6th, 2023.

[7] https://ruffle.rs/. Ruffle. Accessed March 6th, 2023.

[8] https://archive.org/details/Firefox_Chrome_Adobe_Flash. Internet Archive. "Firefox and Google Chrome with Flash Player." Accessed (for this paper) March 6th, 2023.

[9] https://archive.org/. Internet Archive. Accessed March 6th, 2023; https://www.webarchive.org.uk/en/ukwa/. The UK Web Archive. Accessed March 7th, 2023; https://svn.apache.org/. The Apache Software Foundation. Accessed March 7th, 2023.

[10] https://www.nationalarchives.gov.uk/PRONOM/Default.aspx . PRONOM. Accessed February 28th, 2023.

[11] https://wikidp.org/search?q=Adobe%20Flash. WikiData for Digital Preservation. "Adobe Flash." Accessed February 28th, 2023.

[12] https://coptr.digipres.org/index.php/File_Formats_and_Metadata_Formats. Community Owned digital Preservation Tool Registry (COPTR). "File Formats and Metadata Formats." Accessed March 7th, 2023.

[13] P. A. Senster. "The design and implementation of Google Swiffy: A Flash to HTML5 converter." Master's thesis, 2012: oai:tudelft.nl:uuid:cab4b862-d662-432a-afa4-45ccb725177f.

[14] Yogesh Maheshwari and Y. Raghu Reddy. 2017. A study on Migrating Flash files to HTML5/JavaScript. In Proceedings of the 10th Innovations in Software Engineering Conference (ISEC '17). Association for Computing Machinery, New York, NY, USA, 112–116. ¡https://doi.org/10.1145/3021460.3021472 ; https://www.hurix.com/convert-flash-based-websites-

html5/ and https://pixelplex.io/blog/tools-to-convert-flash-to-html5/ are examples of websites that offer more information on migration concerning Flash objects.
[15] Anna Mladentseva (2022) Responding to obsolescence in Flash-based net art: a case study on migrating Sinae Kim's Genesis, Journal of the Institute of Conservation, 45:1, 52-68, DOI: 10.1080/19455224.2021.2007412.
[16] https://ruffle.rs/#usage. Ruffle. "Usage." Accessed March 7th, 2023.
[17] https://github.com/webrecorder/pywb. Github. "pywb." Accessed June 25th, 2023.
[18] A video from the International Internet Preservation Consortium's 2021 Web Archiving Conference (IIPC WAC) demonstrates how this works: https://www.youtube.com/watch?v=XvBt31KgeSk. Youtube. "Not Gone in a Flash: Keeping Flash Accessible in Web Archives (IIPC WAC 2021 Presentation)." Accessed March 7th, 2023.
[19] The source code for this solution is part of the Webrecorder repository on Github. https://github.com/webrecorder/pywb-remote-browsers. GitHub. "pywb-remote-browsers." Accessed March 7th, 2023.

Software Preservation After the Internet

Dragan Espenschied
Rhizome
USA/Germany
dragan.espenschied@rhizome.org
0000-0003-1968-6172

Klaus Rechert
University of Applied Sciences Kehl
Germany
rechert@hs-kehl.de
0000-0002-2454-4374

Abstract – Software preservation must consider knowledge management as a key challenge. We suggest a conceptualization of software preservation approaches that are available at different stages of the software lifecycle and can support memory institutions to assess the current state of software items in their collection, the capabilities of their infrastructure, and completeness and applicability of knowledge that is required to successfully steward the collection.

Keywords – software preservation, knowledge management

Conference Topics – Sustainability; From Theory to Practice

I. Software preservation as knowledge management

This article considers software preservation as providing *continuous access* to *reproduced performance*: Different from software in active use that exposes a lot of touch points to larger systems and processes, preserved software is kept available in a historicized and more constrained archival context. This (idealized) setting allows to trace and comprehend the capabilities of all kinds of legacy software objects into the future.

Software objects are regarded as having a workable boundary definition, including blurry objects with parts of their resources or performance located remotely [1]. A boundary definition typically applies to software that is in some sense "unique" from a particular point of view. For instance, from the perspective of a data science research project, computational processes developed for the project need to be reproducible; for a museum of digital art, artworks in the collection that were collected at different points in history need to have their performance available for exhibition and research; a memory institution concerned with digital work environments will want to make available legacy productivity software like word processors. In any of these cases the software object in focus can be composed of multiple artifacts including large amounts of adjacent software and dependencies that are out of the preserving party's reach or control, or might be logistically impossible to turn into local artifacts—that's the "outside world." The software object's boundary thus has to be defined in terms of its performance capabilities at a certain point in time.

We're further defining the performance of software and the reproduction of that performance as a continuum that is in sync with the lifecycle of a software object as it moves from being actively developed, then maintained, and finally encapsulated.

We suggest a conceptualization of software preservation approaches that are available at different stages of the software lifecycle and can support memory institutions to assess the current state of software items in their collection, the capabilities of their infrastructure, and completeness and applicability of knowledge that is required to successfully steward the collection. Ideally this conceptualization can serve as a guide for improving the understanding of the complexity of software preservation: Software in its different manifestations—as source code, installable binary, installed / configured binary, and remote process—, its different versions—each potentially exposing different characteristics, e.g. adaptations for different markets, languages, and user groups—create a high-dimensional space that is difficult to oversee and makes it hard to navigate to-

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

wards formulating desired preservation goals. However, as these goals become more defined, gaps in preservation knowledge and capabilities can be identified and addressed with new research and infrastructure building projects.

II. How software is made and preserved

A software object that performs and a software object that has its performance reproduced are identical on the artifact level: both the item in focus (such as a particular executable) and the software environment it is embedded in are identical in both stages, bit for bit. The assessment differs according to the activities required to produce or reproduce the performance—the care work that supports a software object—and the level of connectedness of the object to the world outside its object boundaries.

In the continuum from performance to reproduction of performance, three stages can typically be observed. The activities defining each stage are also available for preservation and are structurally based on different utopias: ideas about what will be done to the software object in the future in the service of preservation.

A. Active Development

When a piece of software is under active development, it is tightly connected to and dependent on the outside world. Through constant modification, which might expand or otherwise change the software's capabilities, interactions with other software in an ever-changing environment are kept intact. No matter if programmers aim to produce discrete versions or follow a rolling release model, this stage contains a whole additional level of performance: the one required to build the software. Only if the build performance succeeds will the actual desired performance of the software become available.

Preservation at this stage is based on the utopia that programmers will work on and constantly adapt the software to keep its tight integration with the outside world functional. This approach offers the greatest flexibility: over time, a software could be transformed from a desktop into a mobile application, take advantage of new kinds of displays and input devices, be connected to the latest data sources, etc., thereby matching expectations of regular users towards regular contemporary software.

The knowledge required to keep active development going is large and not static: as changes in the outside world happen, some knowledge about outdated components will become obsolete while new knowledge about updated components will need to be integrated into the software development process. This suggests that a history of versions of the software object will be too difficult to keep continuously accessible, unless the capacity for preservation grows with every version created. With the software object being changed continuously it can be expected that knowledge of how to operate and evaluate it will need to be adapted as well.

The potential for knowledge sharing among memory institutions or preservation practitioners is very low at this stage, as all knowledge is object-specific, and the activities rather demand an immersion into software development communities.

Versions of the software object created during active development might be used in the maintenance stage.

B. Maintenance

When a software object is maintained rather than actively developed, development activities are reduced and often focused on adjacent tools and patches. Instead of running the whole build process for an object to make it perform in changing environments, small fixes are applied, operating system settings and driver configuration options are tweaked, and possibly other software tools that improve compatibility with legacy software are used to expand the lifespan of the object. A single version or multiple versions of a software could be created during the active development stage with the plan to later maintain them.

Preservation at this stage is based on the utopia that there will be some clever trick available in the future that allows for a legacy software object to be performed. Patches and tweaks need to be developed or existing tools repurposed to account for a static software object being embedded in a highly dynamic environment.

The capacity required in this stage is significantly smaller than that for active development. Specific knowledge about how to operate the software object is much less dynamic than during active develop-

ment, because no new versions of the object in question are produced. Knowledge about how to interface existing versions with the changing outside world remains not static, just like in active development new information will have to replace obsolete information. However, there is potential for that knowledge to be generalizable. It is likely that certain classes of objects that can benefit from the same tweaks will be identified. For instance, software that requires a CD-ROM drive can be set up with a virtual CD-ROM driver, Adobe Flash software might be made accessible using the ruffle library, etc.

Over time, the software object will gradually lose its connection to the outside world, as configuration options become unavailable with new versions of operating systems, drivers, and utilities, and required tools will appear too difficult to further develop. At some point the software object will become impossible to perform, perhaps with the last resort being legacy hardware running contemporaneous systems.

Knowledge and tools collected in the maintenance phase might be used in the encapsulation phase.

C. Encapsulation

Encapsulation is an option made possible by emulation [2] and dedicated software preservation frameworks [3]. A fixed version of a software artifact, plus all its dependencies, adjacent tools, and external resources are packaged as immutable disk images, file systems, web archives, etc. and performed by an emulator or a set of emulators orchestrated in a simulated network environment.

All interactions with the outside world happen via managed interfaces of a software preservation framework that controls the emulator or set of emulators. For instance, graphics and sound emitted from an emulator are captured and exposed to the outside world, signals from input devices are translated by the preservation framework into signals that are understood by the emulator.

The utopia of encapsulation is that there will always be emulators in the future and software preservation frameworks will continue to be actively developed.

Object specific knowledge is minimized to only be concerned with how to operate the encapsulated software object. Since this object is not supposed to ever change, and will always perform in the same environment, this knowledge is static. What changes over time are emulators and preservation frameworks, which will need to always accommodate current technical architectures and platforms. Any future issues with the reproduced performance of preserved software objects are to be solved on a framework level. This dynamic knowledge about the preservation framework is highly generalizable, ideally the same for all possible objects, and can be widely shared with a large number of peers with different specializations.

III. Improving preservation capacity

Each stage of a software object is described above in ideal and abstract terms. Especially in a preservation context it is quite unlikely that any of the stages will be observed in their pure form, and mixtures are to be expected, depending on the complexity, size and connectedness of the software in question and its setup.

When framing software preservation as activities enacted on software objects, thus as a knowledge management challenge, it becomes necessary to radically reduce the actively available knowledge required to reproduce the performance of software.

A decision that a software object should be preserved usually happens at a time when it doesn't make sense anymore for the original person or team doing the active development to continue that activity. For instance, software objects produced during an artist residency, or a research grant will need to be taken care of when these projects conclude, and the personnel involved need to move on to other projects.

Institutions as well as communities won't be able to collect, connect, and find ways to apply an ever-growing body of information on software over time. As long as it can be assumed that certain desirable properties of current networked computer systems should persist into the future—that more or less arbitrary parties can participate in and help develop the overall software environment with at least some degree of autonomy—the world outside of a software object will always keep changing. Hence each

new class of software being collected and preserved bears the risk of requiring significant amounts of previously unmanaged knowledge. Yet there is only so much documentation a person can read, or a community can uphold as practice.

Looking at just a single software object in isolation, a care approach modeled after active development makes sense. However, within a collection or archive, it imposes a limit on the number of software objects that can receive preservation treatment and on the time this activity can be sustained. The more knowledge is generalizable instead of object specific, the more institutions and communities can support each other and pool their resources to improve preservation capacity for the field as a whole. Hence, the closer software preservation can move objects towards the encapsulation stage, for single items, for collections, and for software overall, the more likely future generations will be able to explore a rich, diverse, and equitable history of software.

IV. Software preservation realism

It is true that spectacular restoration projects were realized working with legacy source code. [4] [5] They also make for exciting stories as typically important figures from the history of computing and specialized communities, often from the enthusiast space, are involved. Yet exactly these inspiring stories should be interpreted as indicators of the risks of relying on active development for software preservation. While it might be possible to recruit highly skilled developers and knowledgeable hobbyists to work on a groundbreaking software object like an influential game or a landmark operating system—alternatively, pay them well enough to do so—, this is unlikely to happen for an under-appreciated artwork, custom research software, or, plainly boring yet essential software as developed for administrative purposes in government and commerce.

Preservation projects focused on active development are also more likely to succeed for the "classic" model of software creating in which programmers work with local source code and locally available libraries to produce a whole piece of software or component to be packaged and shipped via carrier media or a network connection.

Software Development After the Internet works quite differently. Distributed package managers, interpreted computer languages, and "continuous integration" build processes dominate mainstream development practice and afford developers with previously unknown levels of nimbleness and powerful abstractions. Here the build process has become a performance before the performance, with likely as many variables to consider as for the performance of the software object that is being built. Unless it can be demonstrated that the build process actually works, it is not even possible to assess if all the required dependencies and external resources are available in some form. Since packages and libraries from remote repositories can also change without notice, it becomes increasingly difficult to even deliberately delineate versions of the software object that uses them, and temporarily suspending continuous care for a software object runs the risk of opening a knowledge gap that might turn out to be impossible to close later.

Given these considerations it seems reasonable to move out of the active development stage for preservation purposes and only deal with challenges of an object's "main performance" that is available after the build. This even makes sense when considering that there is no technical difference between reproducing a build performance or reproducing any other software performance, meaning that a build process could be moved to the encapsulation stage just like its final product. The practicality of this approach has to be decided on a case-by-case basis. For instance, if a software object requires rebuilding to produce different desired results in its main performance, the build performance could be made reproducible as well. The usual restrictions of encapsulation would apply, in particular the loss of unmediated interaction with the outside world.

In some cases, it might also not be possible to leave active development behind, in particular when the tools and techniques used to build the software are not well understood or difficult to control due to their novelty or because they're highly proprietary and opaque. Keeping active development going for long enough to gather sufficient knowledge to move to the maintenance or encapsulation stage could be the only way to develop a long-term perspective for certain types of software objects, such as games built

with proprietary toolkits, software requiring highly secured proprietary online accounts, or access to proprietary data.

Entering the maintenance stage is attractive when active development is uneconomical, and Shared knowledge can be utilized. Despite the software industry's push to move all users into subscriptions for most products, there is a wealth of knowledge around on how to keep legacy software running and operational past official support times, by tweaking aspects of new systems to cater for the needs of legacy software objects. Sometimes legacy hardware computer setups are available, or systems are deliberately disconnected from the internet to prevent any unintended automatic updates. Many of these tweaks can be abstracted and applied to several objects of a similar technical composition. Overall, maintenance moves the attention of programmers outside of the object to be preserved to the environments it should perform in.

This approach is for the most part offering the same performance and performance quality as active development—a freshly built object will be as responsive and snappy to interact with as a maintained one that is running on a similar system—, yet at some point the connections to the outside world will become impossible to keep going and the object's performance will degrade.

Returning from maintenance to active development with the plan to update the software object once and to then resume with maintenance can be very expensive and risky. Since active development was suspended while the object was maintained, a large knowledge gap might have appeared that needs to be bridged before development can start. It is also hard to predict for how long the result of such a one-time fix will be able to reproduce the desired performance. An update produced with significant effort might become outdated pretty quickly, calling for another potentially expensive active development phase.

The encapsulation stage in many cases relies on products of active development and maintenance: a software object needs to be built to exist in the first place, and maintenance knowledge can be used to construct a suitable environment to package alongside. Once that is done, knowledge can quite cleanly be separated into static object specific and generalizable infrastructure knowledge. Future risk is reduced to the need for suitable emulators being available and the maintenance of emulators' interfaces with the outside world on the framework level. This means the framework level is ideal for collaboration and most knowledge and development effort can be shouldered in concert by otherwise not affiliated actors. Missing features in an emulator or preservation framework can be identified by practitioners, and stakeholder groups or open fundraising efforts can then commission work to the benefit of any software preservation use case. In an ideal world, emulators and preservation frameworks would be the only places that require active development in order to provide continuous access.

Of course, encapsulation in reality has some drawbacks. Software objects that use bleeding edge or proprietary devices and components or data sources will typically not be possible to capture in full right after creation. For instance, at the time of writing, this is true for projects using virtual reality or augmented reality, dealing with software embedded in a highly competitive market with constantly changing devices, development environments, and real-time online services. Ongoing research on singular objects and classes of objects under active development or maintenance is required to understand which features need to be included into emulators and preservation frameworks.

Additionally, accessing software performance via emulators and preservation frameworks will never be as direct as using software under active development, and to some degree, under maintenance. There will always be some layer users need to pass to access a reproduced performance versus a regular performance, because emulators will have to be spun up and configured by the preservation framework to fit the presented object, and noticeable differences in usage conventions and visual design will contribute to an impression of media discontinuity. This means that any encapsulated software object is necessarily historicized.

V. Overcoming Long-Term Limits on Managing Software Knowledge

Extensive documentation on legacy software products is available in the form of printed and electronic books in libraries. Additionally, the preservation community active on the web provides us with a wide variety of internet artifacts containing tips and hints on configuration, usage and, most importantly, repairing non-functional software products. While this wealth of documentation is necessary and useful, as time goes by, it will become less actionable. The information available was prepared for contemporaneous users and omitted lots of knowledge regarded as implicit in its time. This concerns in many cases basic instructions on how to configure and operate systems that are now deemed obsolete and have fallen out of use. As every change in the software landscape potentially adds another layer of knowledge, demanding preservation professionals to make themselves familiar with everything they need to fully understand any software they are supposed to preserve, is unrealistic and ethically questionable. Similarly, preservation professionals should reflect on their reliance on enthusiast communities and creators for keeping knowledge active and easily retrievable.

Fully configured, encapsulated computing environments (in most cases in the form of a disk image combined with instructions on how to connect and start it up in an emulator) already can serve as a technical embodiment of knowledge, as they can be used without having to look up how to construct one from scratch.

As a next step, recordings of knowledgeable users interacting with encapsulated systems inside a preservation framework can be made so they become deterministically replayable in the future [6]. These recordings can be used to automate simple, recurring tasks, for instance, configuring applications or an operating system; a particularly important use-case for software preservation is automated installations of applications requiring user input during setup. Furthermore, these recordings—together with user annotations—can act as executable documentation, allowing users to follow operational steps and if necessary, take over and adapt a recording to similar tasks, which could then be annotated and stored as a new automated task in a library. Users could choose to have these automations executed in the background, for instance on many encapsulated environments that need to be reconfigured in a similar way or watch the execution to learn the steps.

Even though such recordings act as actionable, executable documentation, a potential library of such recordings will quickly grow into a silo that's difficult to maintain, containing highly context sensitive information for which no concept for indexing apart from manual annotations and basic technical metadata currently exists. Over time, it will become highly desirable to interact with legacy systems on an increasingly abstract level. For instance, to not have to learn how to load a file in dozens of different applications that might run on top of a bunch of operating systems with differing user interface conventions, recordings would need to become much more variable than fully deterministic.

Feeding existing recordings to a learning algorithm has the potential to make this abstraction possible, taking advantage of the similarities in user interface design conventions used within certain time periods. If a sophisticated enough model that matches semantically described desired activity and user actions can be created, it might be trained on legacy tutorial screen capture videos released by software vendors or created by user communities as released on YouTube or similar public video sharing platforms.

1. REFERENCES

[1] D. Espenschied and K. Rechert, "Fencing Apparently Infinite Objects," in *Proceedings of the 15th International Conference on Preservation of Digital Objects*, Boston, U.S., 2018.

[2] D. S. Rosenthal, "Emulation & virtualization as preservation strategies," 2015.

[3] E. Cochrane, K. Rechert, J. Oberhauser, S. Anderson, C. Fox, and E. Gates, "Useable Software Forever," *IPres 2022 Glasg. 12—16 Sept. 2022 Www Ipres2022 Scot*.

[4] G. Mastrapa, "The Geeks Who Saved Prince of Persia's Source Code From Digital Death," *Wired*, Mar. 10, 2023. Accessed: Mar. 10, 2023. [Online]. Available: https://www.wired.com/2012/04/prince-of-persia-source-code/

[5] "The ReCode Project — Matthew Epler," Mar. 10, 2023. https://mepler.com/The-ReCode-Project (accessed Mar. 10, 2023).

[6] J. Oberhauser, R. Gieschke, and K. Rechert, "Automation is Documentation: Functional Documentation of Human-Machine Interaction for Future Software Reuse," *Int. J. Digit. Curation*, vol. 17, no. 1, Art. no. 1, Sep. 2022, doi: 10.2218/ijdc.v17i1.836.

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

PUBLISHING AGRICULTURAL DATA FROM THE MORROW PLOTS

The Value and Logistics of Preserving a Long-Term Research Experiment

Bethany G. Anderson
University of Illinois
United States
bgandrsn@illinois.edu
0000-0001-6602-1312

Sandi L. Caldrone
University of Illinois
United States
caldron2@illinois.edu
0000-0001-6392-5279

Joshua K. Henry
University of Illinois
United States
jkhenry@illinois.edu
0000-0002-7826-5960

Heidi J. Imker
University of Illinois
United States
imker@illinois.edu
0000-0003-4748-7453

Andrew J. Margenot
University of Illinois
United States
margenot@illinois.edu
0000-0003-0185-8650

Sarah C. Williams
University of Illinois
United States
scwillms@illinois.edu
0000-0001-7968-1870

Abstract – The Morrow Plots at the University of Illinois Urbana-Champaign are the longest-running continuous experimental agricultural fields in the Americas. At iPres 2022 we reported on work to curate, preserve, and visualize planting, treatment, and yield data collected from the plots' nearly 150-year history. This paper provides an update on these efforts over the past year and, with special emphasis on the data's scientific and cultural value, discusses the importance of collaborative and interdisciplinary work within the Morrow Plots stakeholder community to publish the dataset, and identify necessary next steps.

Keywords – data, agriculture, archives, curation, collaboration

Conference Topics – We're all in this together

I. INTRODUCTION

The Morrow Plots, located at the University of Illinois Urbana-Champaign, are a set of well-known experimental agricultural fields noted for both their scientific and cultural importance. The plots are the site of a long-term research experiment (LTRE) to test the effects of crop rotation and were established in 1876, making them nearly as old as the university itself. The plots are of such significance that the university's College of Agricultural, Consumer and Environmental Sciences (ACES) is planning a sesquicentennial event in 2026. In preparation for this celebration, the authors became interested in enabling greater access to various materials pertaining to the plots, including the data resulting from this LTRE right in the heart of our campus.

Various kinds of experimental data have been collected from the plots in their long history, but there had not been an attempt to consolidate the data into a single, cohesive, well-documented dataset that could be publicly shared and used by others. At iPres 2022, we reported on our early efforts to establish the "Morrow Plots Data Curation Working Group," a cross-unit collaboration involving the College of ACES, the University Library, and the University Archives [1]. In this paper we describe an update on the group's progress since last year including the recent assembly and publication of a planting, treatment, and yield dataset, which was made possible by blending data and preservation expertise with deep disciplinary engagement and knowledge. We describe the various stakeholders involved, their interests in this project, and the data release process, including our efforts to engage stakeholders throughout.

A. A Brief History of Change

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

In 1876, Professor Manly Miles broke ground on "Rotation Experiment 23" which would later become known as "The Morrow Plots" to test growing conditions for corn, Illinois' most important crop. Every year since, agricultural researchers have planted these plots, located just off the Main Quad, with a combination of corn and other common local crops like oats and clover. Plot divisions allow for comparisons between soil fertility treatments and crop rotation schedules. Over time, the plots have been divided and subdivided as new treatments (for fertility input specifically) were introduced. Although the experiment has continued uninterrupted for over a century, change has been a constant from the beginning.

Shortly after the launch of the experiment, Professor Miles left his post at the University of Illinois. George Espy Morrow, the experiment's namesake, then assumed responsibility in the fall of 1876, only a few months after the beginning of the work earlier that spring. While little is known about the details of this first hand-off, it was by all accounts swift with the termination of Manly Miles' contract in June of 1876 [2] [3]. This turnover of the experiment's leadership would be the first of many in its nearly 150-year history. Since the formation of the Morrow Plots Data Curation Working Group in 2018, stewardship of the plots has changed hands between three different parties. The Morrow Plots are now overseen by the laboratory of Professor Andrew Margenot, current chair of the Morrow Plots Steering Committee (and a co-author of this article).

Each time the Morrow Plots experiment transitioned to a new steward, there was much to be considered as a part of the switch. From the fundamental understanding of the details of the experiment and the nature of its importance, to the recordkeeping and data storage practices employed to ensure the longevity of the experiment's value. These challenges have been repeatedly brought home to us as members of the Morrow Plots Data Curation Working Group. In addition to changes in plot management over time, within the relatively short history of this working group, there have been major changes to administration in our respective colleges as well as changes in working group membership itself as people transitioned on or off the working group in keeping with different roles or jobs. Thus, we recognized early on that stakeholder awareness and engagement would require perpetual attention.

II. Stakeholders

Given the Morrow Plots' prominence (physically, historically, and culturally), multiple units at the University of Illinois have an interest in the plots, their history, and their associated data (Fig. 1); in particular, the College of ACES and its Department of Crop Sciences, as well as the Library, the Archives, the Funk ACES Library, and the Research Data Service. Individuals within each of these units also have their own lens for considering the value of anything associated with the plots. For example, researchers from Crop Sciences take an active interest in using the plots for research purposes [4], while archivists are interested in ensuring that this important landmark continues to be represented in the history of the university and that data from the plots are preserved and made available for historical and scientific research use [5] [1]. For life sciences librarians, the Morrow Plots are an important facet of collection management, instruction, and research assistance at the University of Illinois [1].

Additionally, communications groups at the college, library, and even university-level are interested in the plots for their ability to demonstrate the value of agricultural research, convey the impact of the university's history, and connect with the public, including hundreds of thousands of alumni, who remember the plots as one of the university's most famous landmarks. Preserving evidence and data from the Morrow Plots and sharing these materials with local communities and the general public is critical to the University of Illinois' role as a land-grant institution.[1] The plots are important not only for agricultural and university history [6], but also for the advancement of agriculture in the state of Illinois and beyond [7]. Bearing in mind this broad array of stakeholders is important when curating and creating access to the plots' data and engaging

[1] Land-grant institutions are public colleges established by the Morrill Act of 1862, the first federal investment in higher education in the U.S.: https://www.archives.gov/milestone-documents/morrill-act.

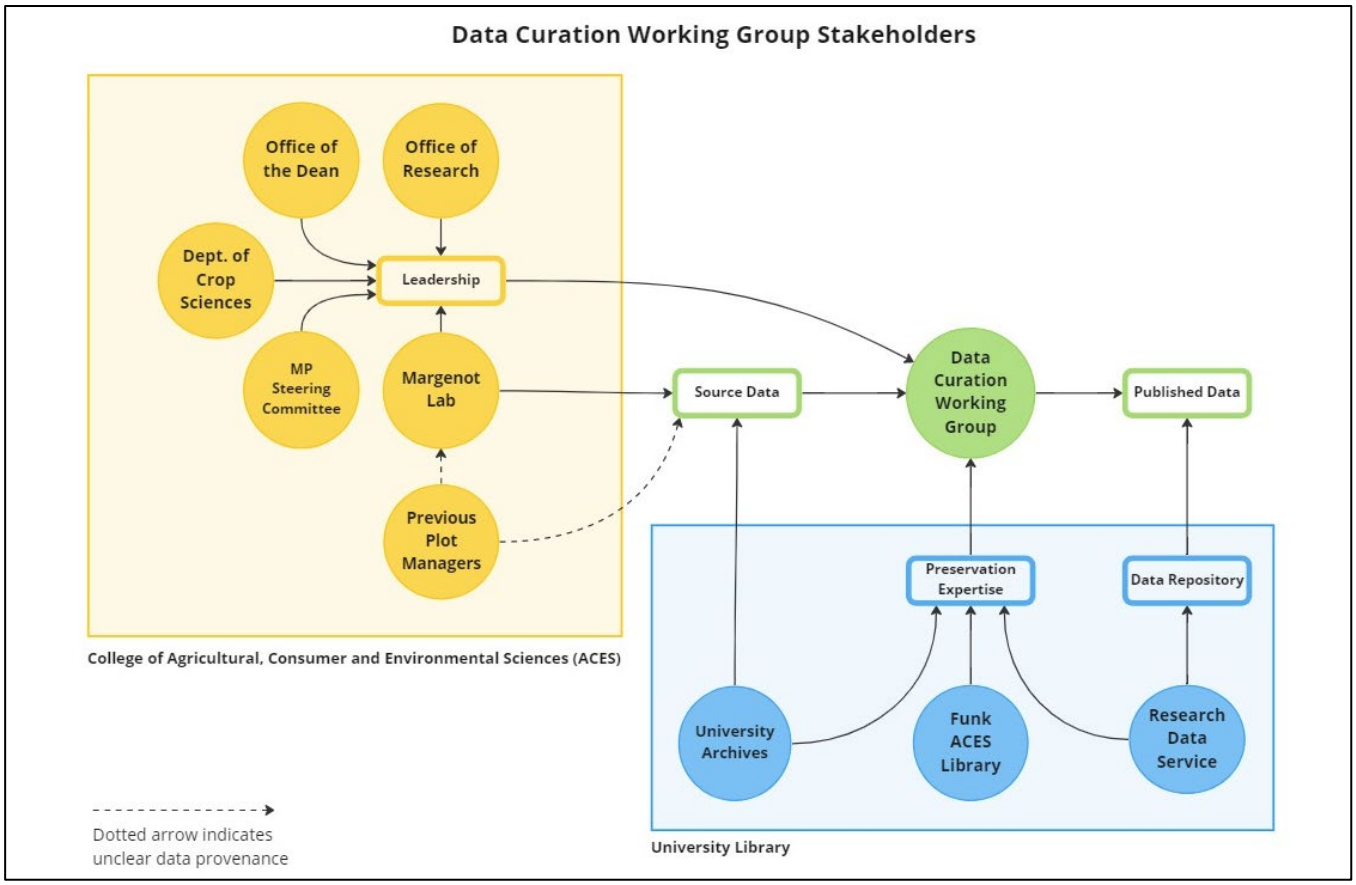


Fig. 1, Data Curation Working Group Stakeholders diagram.

university and external communities in the history and results of this long-term agricultural experiment.

III. Value Perspectives

A. Agricultural Research

LTREs are few and far between in the agricultural world. These are not necessarily synonymous with agricultural experiments, as there can be LTREs that are ecological in nature (e.g., carbon-enrichment experiments), as evidenced by the National Ecological Observatory Network (NEON) funded by the U.S. National Science Foundation [8]. The U.S. Department of Agriculture funds an 18-site network of long-term agroecosystem research experiments [9], established at the turn of the twentieth century after a crescendo of calls for well-designed, at-scale and intentional long-term evaluation of agroecosystems (LTAR) in the U.S. [10] [11]. However, these federally funded twenty-first century efforts are distinct from and relatively much younger than historical LTREs such as the Morrow Plots. A cluster of these historical LTREs were established with the advent of the land-grant institutions in the late 1800s, several of which continue today as some of the oldest LTREs in the world. In addition to the Morrow Plots – the eldest sibling (1876) - there are the Sanborn Fields at University of Missouri (1888), Magruder Plots at Oklahoma State University (1892), and the Cullars Rotation at Auburn University (1911) [12].

LTREs offer unique insights to the sustainability of agriculture. For example, the oldest continuous agricultural experiment in Rothamsted, United Kingdom, has produced multiple and invaluable insights to the effects of agricultural management practices on soil functions and crop productivity [13]. LTREs offer direct observations of how soils change at timescales beyond typical funding cycles (e.g., 5 years), thereby offering insight to sustainability – inherently a timescale function – at scales not usually assessed. In the case of centennial-scale LTREs, the information gained spans the careers of multiple scientists. Emergent agroecosystem processes and properties can also be captured by LTREs, because some functions of cropping systems only manifest at multidecadal timescales. For example, changes in

soil organic matter or yield stability emerge at timescales that exceed most experimental durations of 5-10 years [14] [15]. Finally, the longitudinal data offered by LTREs, especially in conjunction with auxiliary data (e.g., weather), enables model calibration and validation at a scale that enables backcasting and forecasting at longer time ranges and with higher confidence.

Some have questioned the utility of LTREs such as the Morrow Plots [16] and even Rothamsted [17]. Classic arguments for the constraints of LTREs include:

- Small plot size: as a result of subdividing the already relatively small crop rotation plots into fertility input treatments, the 24 distinctly managed plots that currently make up the Morrow Plots are less than 0.01 ha in area. Small plot size means that observations such as yield may be prone to variability or random effects that decrease sensitivity to treatment effects (e.g., drought) or make a given plot more susceptible to data loss (e.g., rodent damage).
- Replication: preferably randomized, replication is key to enable statistical analyses of response variables. However, the Morrow Plots do not have strict replication of treatments, precluding analysis of variance (ANOVA) for replicated block or complete block designs [16]. On the other hand, longitudinal studies are still enabled [18], as well as approaches such as exploratory factor analyses and multivariate analyses that enable detecting signals in non-replicated treatment plots over time [19].
- Context-dependence: from soil type to climate conditions to geographic region, the insights of a given LTRE will be limited in inference space. This is an issue for any field experiment, which is why a network of LTREs such as the USDA-funded LTAR is essential [9].
- Changes in treatments over time: as with any long-term experiment, changes must be made to treatments to keep pace with current practices, including: crop cultivars (e.g., modern hybrid) or even species (e.g., Morrow Plots switched rotation from oats to soybean to reflect recession of animal draft power by mid-1900s), fertilization approaches (e.g., rates, sources), planting densities, tillage, and pest management. However, strict adherence to the original treatments of any long-term experiment will rapidly make the experiment obsolete (e.g., use of bone meal as a fertility source and open-pollinated maize varieties in the Morrow Plots). For this reason, LTREs around the world have kept pace with changes in specific agricultural practices while maintaining the overall concept of a treatment (e.g., "high fertility input" treatments changing sources and rates over time). In fact, updating treatments to reflect contemporary changes helps identify how advancements in agricultural technologies such as hybrids or fertilization have impacted yields and agroecosystem properties such as soil organic matter content [16] [18] – serving as a living record of such changes.

As with any experiment, there are limitations that must be gauged with benefits for scientific insights. Despite the general and specific limitations, the Morrow Plots still offer unparalleled insights to the long-term impacts of crop rotation and fertility management by virtue of its sesquicentennial duration and the archiving of soils sampled since 1904, including soil organic matter [16], microbiological community composition [19] [20], yields [16] [18] and even soil formation [21] and mineralogy [22] [23]. Additionally, the experiment and its soil archive enable evaluation of non-agricultural biogeochemical processes, such as lead deposition from coal combustion in the early 1900s [24]. A comprehensive description of layout and history of the treatments, site conditions, and yields have been extensively reported elsewhere [18].

Beyond these trade-offs in scientific value, logistical constraints to LTREs include the resources needed to continue experiments. Here, the Morrow Plots are unique in that they have enjoyed strong support from the University of Illinois, no doubt in part from their high visibility as a centrally located and national historical monument site as of 1968 [6] on the campus of the state land-grant university. The relatively small size of the Morrow Plots – a key disadvantage for its research applications – also incurs a lower maintenance cost. In contrast, a large-

scale LTRE such as the University of California Davis Century Experiment (e.g., 1 acre replicate treatments) has high costs of operation, which contributed to the early demise of this hundred-year LTRE in year thirty-three [25].

B. College of Agricultural, Consumer and Environmental Sciences

The University of Illinois has a long history of agricultural education and research, especially related to agricultural experiments. An Agriculture Department was among the university's initial units, and a College of Agriculture was established ten years later in 1877 [26]. Following the Hatch Act of 1887, an Agricultural Experiment Station was established at the University of Illinois to oversee the management of several ongoing and new experiments [27]. Originally known as "Experiment 23," the Morrow Plots was one of these experiments but focused specifically on the effects of crop rotation [28]. Today, it is the second longest-running agricultural experiment in the world, following experiments at Rothamsted, but to the University of Illinois it also holds relevance as a long-term experiment that has roots in ACES's early history and the founding of the university as a land-grant institution.

A search of scholarly outputs that either name or cite research from the Morrow Plots since 2019 returns over one hundred entries, and in varied disciplines from soil science and agronomy to journalism and poetry. The results are not just U.S. or North America centric. Citations and mentions from Europe, South America, East Asia, the Middle East and Africa are also represented. This international scholarly and scientific profile holds significant importance to the University of Illinois and the College of ACES. Because of this reach, the preservation and presentation of the outputs from this LTRE are extremely valuable. In addition to the scholarly impact, the Morrow Plots hold a specific historical and cultural significance to the university community. The "hope that hunger and privation are not the inevitable fate of man," noted in the proceedings of the plots' initiation as a registered National Historic Landmark in 1968 underlines not only the scientific value of the experiment but the aspiration that through scientific research and collaboration we can best the worst challenges of our time [29].

C. University Archives

As the official repository for the University of Illinois, the University Archives has a mandate to broadly document the history and activities of the university. The Archives aims to preserve and make available for research use records and personal papers that capture various facets of life on campus including students' experiences, the work and contributions of faculty, alumni, and staff, and decisions of university administration. As a part of this mandate, the Archives has routinely transferred and preserved records from the University of Illinois' science units and programs. A significant part of the University of Illinois' early history and curricula was focused on agriculture [30], which is due in part to being one of the land-grant institutions that came out of the Morrill Land Grant Acts in the nineteenth century [31].

The University Archives maintains records that shed light on the university's agricultural history as well as the plots' long history as an initial experiment, but also its changing hands over time (and university politics surrounding its significance). The vast majority of the Archives' materials pertaining to the plots are administrative records, and the Archives has a dearth of materials documenting its scientific significance, especially in the form of historical data.[2] In 2019, however, the Department of Crop Sciences transferred a notebook to the Archives containing rotation and yield data spanning from 1876 through 1913, which was subsequently digitized [32]. While this notebook is a significant acquisition that fills this gap in the Archives' holdings, and the Morrow Plots Data Curation Working Group has identified overlapping and later yield data [1], we hope to locate, preserve, and make available other extant data for both historical and scientific research use.[3]

[2] See the Morrow Plots LibGuide for a list of records from the University Archives, https://guides.library.illinois.edu/Natural_and_Applied_Sciences_Archives/Morrow_Plots.

[3] The University Archives has photographic evidence of other notebooks which exist that possibly have additional data. See Agriculture, Consumer, and Environmental Sciences Photograph File (Born Digital Records), Record Series 8/1/57,

Preserving and making available data from the Morrow Plots helps make the case for access to historical data more generally, especially given the long engagement from the scientific community with the plots (as noted above). There are myriad reasons why ensuring broad and public access to historical data (especially longitudinal data) has great benefit to current science as well as to other stakeholders who have historical, administrative, and general interests including research on climate change and ecological shifts over time (e.g., Climate Data Online). Increasingly, researchers are seeking to identify and reuse historical (analog) data, but finding such data in the first place may be difficult [33]. One challenge is that it is not immediately evident where researchers can look for historical data. For example, academic archives in particular have not been actively engaged in curating and preserving data, let alone historical data [34]. The efforts of the Morrow Plots Data Curation Working Group speak to the need to create access to and ensure the archival preservation of historical data, but also underscore the ways in which such a project benefits from collaboration and engagement with a multitude of stakeholders. From an archival perspective, this helps emphasize the need for more historical data to be identified, transferred to an archive or a disciplinary repository, and thoughtfully and carefully curated for public use. Archives have a vital role in this space, especially bearing in mind the value of historical data when appraising scientific records while also enhancing access to historical data already held in their collections. At a land-grant institution like the University of Illinois, ensuring public access to such data helps foster greater engagement with the university's scientific heritage while recognizing other potential and future stakeholders across the State of Illinois and beyond.

D. Research Data Service

As one of the stakeholders in the Morrow Plots Data Curation Working Group, the Research Data Service (RDS) provides campus researchers with the expertise, resources, and infrastructure necessary to responsibly manage and steward data. Housed within the Library and serving the entire campus research community, the RDS provides workshops and guest lectures on data management best practices, reviews the data management plans typically required with grant applications, consults with researchers and labs on policies and procedures, and operates the Illinois Data Bank, an open-access institutional repository for research data and its associated documentation and code.

By aggregating, cleaning, and documenting the plots data, the RDS gained invaluable experience from the process leading up to publication. The experience of curating a historical dataset from a LTRE and navigating the interests of the wide variety of stakeholders involved, prepares us for consultations on datasets and data management plans from other longitudinal projects. This experience can be applied not only within the Illinois research community, but also in the wider Data Curation Network, a national network of data repositories and memory institutions to which the university belongs.

The content of the dataset has clear value to agricultural researchers, but the form of the dataset has value to other researchers as well. In preparing the dataset for publication, great care was taken to clean and format the data according to the Tidy data specifications for improved clarity and interoperability [35]. In the interests of open science, the dataset was also meticulously documented, and the cleaning process recorded both in the codebook published alongside the data and in a GitHub repository containing all of the R code used in the process [36]. The RDS can point to these as a model for data producers in a wide range of fields.

The dataset also lends itself well to educational use. At just over 3,000 rows, the table is easy to open and navigate without special skills or software, and since much of it was manually recorded, it's highly human readable. Additionally, the concepts of planting, fertilization, and yield are easily understandable without background knowledge. Lessons about data management and the associated tools and software require sample datasets. The Morrow Plots data would not only be easy to integrate into the classroom for the reasons listed above, it would also give students opportunities to

University Archives, https://digital.library.illinois.edu/items/81c58e60-57f3-0134-1dc2-0050569601ca-6.

learn not just about data but also about history and preservation. Finally, exposing the data to a wider audience, including students, opens the door to imaginative reuse.

IV. Logistics

A. Publishing the Planting, Treatment, and Yield Dataset

Recent work undertaken by the Morrow Plots Data Curation Working Group included compiling planting, treatment, and yield data covering 1888 to 2021 from three sources: 1) an internal tracking spreadsheet, 2) published yield tables, and 3) an archival notebook. The tracking spreadsheet kept by farm managers provided crop data for 1955 to 2021. As the most detailed of the three sources, the tracking spreadsheet provided the majority of the variables and the basic structure of the compiled dataset. The dataset was extended backward in time with the help of yield tables covering 1888 to 1954 previously published in a 1982 University of Illinois Agricultural Experiment Station Bulletin [21]. The yield table data was highly condensed for print publication, but once teased out, covered many of the same variables in the tracker. It did not, however, include the specific planting dates or the particular crop varieties for those years. As much as possible, these two variables were supplemented by the third source, the handwritten notebook in the University Archives covering 1876 to 1913 [32].

Data from all three sources was imported into RStudio to be cleaned and compiled. Summaries, spot checks, and exploratory data visualizations were used throughout the process to ensure data integrity. The R code used to clean, combine, and format the dataset is publicly available in R Markdown format on GitHub [36]. For more information on the cleaning process, please see our iPres 2022 publication [1]. The working group opted for the Tidy data format for tabular data employing one column per variable and one observation per row [35]. One of the benefits of the Tidy format is that it takes implied information (e.g., color coded text in the Yield variable to designate crop damage) and makes it explicit (e.g., the addition of a separate Damage variable), allowing the dataset to be converted to other formats without data loss. The Tidy format is highly machine readable, making it easier to combine with other datasets. We successfully tested this by linking the compiled data to a fourth dataset, an inventory of soil samples periodically taken from the plots starting in 1904. The result is a TRUE/FALSE variable noting whether a soil sample exists for the corresponding plot and year. The Margenot Laboratory is working to make more information about the soil samples publicly available, and we may be able to provide more detail in future versions of the dataset.

The Illinois Data Bank was a natural fit for publishing the aggregated Morrow Plots dataset where it will be both carefully preserved and freely available to the public. All Illinois Data Bank datasets are assigned a digital object identifier (DOI) registered with DataCite along with additional metadata following the DataCite Metadata Schema for improved discoverability [37]. The completed dataset is published in CSV format along with an extensive codebook [5], making the dataset FAIR (findable, accessible, interoperable, and reusable) [38]. The Illinois Data Bank retains published datasets for a minimum of 5 years and after that as long as the data remains relevant. Given the enduring value of the Morrow Plots experiment and our deliberate decision to share the dataset in preservable formats, we can expect the data to remain in the repository's collection indefinitely.

B. Administrative Approval

Although the Morrow Plots Data Curation Working Group formed with hopes of gathering and sharing data from the Morrow Plots, the first phase was exploratory. It was not clear exactly what outcomes might result, let alone how those outcomes might be distributed and under what conditions. Until viable outcomes could be articulated, it was difficult to imagine, let alone seek approval for, such details. Additionally, there was no formal charge for the group. While the group's work progressed and the aggregate dataset came together, informal conversations were happening with researchers within ACES (via the Morrow Plots Steering Committee) and the then-manager of the plots. Periodic updates were also given to administration in both ACES and the Library. However, an informal conversation is, by definition, unofficial, and once we were ready to release the dataset to the public, it became necessary to make it very clear to leadership exactly what we hoped to do. A misstep at this stage could result in surprised,

possibly alarmed, or even angered administrators. Therefore, we opted for a synchronized strategy to update our respective administrators in ACES and the Library and get approval prior to releasing the data. Since there are multiple draws on an administrator's attention, we made sure our communications were clear, concise, and decision-ready. We waited until we had a completed draft dataset ready for previewing in the Illinois Data Bank so that administrators could evaluate exactly what would be shared. In communications we also highlighted the three trickiest aspects of publishing the dataset: authorship (especially given the many hands that touched the data over the years), licensing, and the potential for negative impacts since, after almost 150 years of data collection, there are gaps in some variables. We encouraged review, offered to meet to discuss, and finally listed the individual administrators being contacted in the other unit so that it was clear who was being consulted. We received swift approval from the Dean of the Libraries while the Dean of ACES first consulted with others in the college. We retained this email correspondence to document the decision.

V. Conclusion

With the initial work of collecting, documenting, publishing, and preserving the Morrow Plots planting, treatment, and yield dataset completed, the working group looks to the future with plans to communicate its availability, develop new resources and relationships for its use as an educational tool, and to celebrate the legacy and impact of the experiment with the stakeholders of the Morrow Plots in a sesquicentennial symposium in 2026. Additionally, efforts are ongoing to uncover and include more data and context materials to enrich the existing data. The working group will also document best practices for curating and preserving Morrow Plots data to ensure long-term stability of the data, especially as the stewardship of the Morrow Plots transitions over time.

Like the initial work, these next steps will require a variety of perspectives and expertise to be successful. No individual or single unit could achieve these outcomes. As working group members, we will continue collaborating and engaging others to further preserve, disseminate and enhance the Morrow Plots data. By coming together, the group members are contributing significantly to preserving and celebrating university history, enabling agricultural research, and improving data education.

1. REFERENCES

[1] B. G. Anderson *et al.*, "Cultivating the Scientific Data of the Morrow Plots: Visualization and Data Curation for a Long-term Agricultural Experiment_20220915," presented at the iPres, Glasgow, Scotland, Oct. 2022. doi: 10.17605/OSF.IO/X8YMS.

[2] C. G. Hopkins, W. G. Eckhardt, and J. E. Readhimer, "Thirty years of crop rotations on the common prairie soil of Illinois," 1908, Accessed: Aug. 08, 2019. [Online]. Available: https://www.ideals.illinois.edu/handle/2142/3031.

[3] R. G. Moores, *Fields of rich toil: the development of the University of Illinois, College of Agriculture*. College of Agriculture, 1970.

[4] H. Spangler, "What are the Morrow Plots?," *Farm Progress - Prairie Farmer*, Feb. 02, 2023. https://www.farmprogress.com/crops/what-are-the-morrow-plots-.

[5] Morrow Plots Data Curation Working Group, "Morrow Plots Treatment and Yield Data." University of Illinois at Urbana-Champaign, 2022. Accessed: Mar. 02, 2023. [Online]. Available: https://doi.org/10.13012/B2IDB-7865141_V1.

[6] College of Agricultural, Consumer and Environmental Sciences, *The Morrow Plots: A Landmark for Agriculture*. University of Illinois Urbana-Champaign. College of Agricultural, Consumer and Environmental Sciences., 2001. [Online]. Available: https://hdl.handle.net/2142/113519.

[7] T. Hmielowski, "The value of long-term data in agricultural systems," *CSA News*, vol. 62, no. 1, pp. 4–7, Jan. 2017, doi: 10.2134/csa2017.62.0101.

[8] S. L. Collins and A. K. Knapp, "NEON Should Be Run by Ecologists for Ecologists," *BioScience*, vol. 69, no. 5, p. 319, May 2019, doi: 10.1093/biosci/biz043.

[9] P. J. A. Kleinman *et al.*, "Advancing the Sustainability of US Agriculture through Long-Term Research," *J. Environ. Qual.*, vol. 47, no. 6, pp. 1412–1425, 2018, doi: 10.2134/jeq2018.05.0171.

[10] G. P. Robertson *et al.*, "Long-term Agricultural Research: A Research, Education, and Extension Imperative," *BioScience*, vol. 58, no. 7, pp. 640–645, Jul. 2008, doi: 10.1641/B580711.

[11] M. R. Walbridge and S. R. Shafer, "A Long-Term Agro-Ecosystem Research (LTAR) Network for Agriculture," *Eng. Technol. Sustain. World*, vol. 20, no. 5, pp. 8–11, 2012.

[12] D. D. R. Jr and D. Markewitz, *Understanding Soil Change: Soil Sustainability Over Millennia, Centuries, and Decades*. Cambridge University Press, 2001.

[13] A. E. Johnston and P. R. Poulton, "The importance of long-term experiments in agriculture: their management to ensure continued crop production and soil fertility; the Rothamsted experience," *Eur. J. Soil Sci.*, vol. 69, no. 1, pp. 113–125, 2018, doi: 10.1111/ejss.12521.

[14] T. M. Bowles *et al.*, "Long-Term Evidence Shows that Crop-Rotation Diversification Increases Agricultural Resilience to Adverse Growing Conditions in North America," *One Earth*, vol. 2, no. 3, pp. 284–293, Mar. 2020, doi: 10.1016/j.oneear.2020.02.007.

[15] A. E. Johnston, P. R. Poulton, K. Coleman, A. J. Macdonald, and R. P. White, "Changes in soil organic matter over 70 years in continuous arable and ley–arable rotations on a sandy loam soil in England," *Eur. J. Soil Sci.*, vol. 68, no. 3, pp. 305–316, 2017, doi: 10.1111/ejss.12415.

[16] E. D. Nafziger and R. E. Dunker, "Soil Organic Carbon Trends Over 100 Years in the Morrow Plots," *Agron. J.*, vol. 103, no. 1, pp. 261–267, 2011, doi: 10.2134/agronj2010.0213s.

[17] D. S. Jenkinson, "The Rothamsted Long-Term Experiments: Are They Still of Use?," *Agron. J.*, vol. 83, no. 1, pp. 2–10, 1991, doi: 10.2134/agronj1991.00021962008300010008x.

[18] S. Aref and M. M. Wander, "Long-Term Trends of Corn Yield and Soil Organic Matter in Different Crop Sequences and Soil Fertility Treatments on the Morrow Plots," *Adv. Agron.*, 1997, doi: 10.1016/S0065-2113(08)60568-4.

[19] S. S. Raglin, C. Soman, Y. Ma, and A. D. Kent, "Long Term Influence of Fertility and Rotation on Soil Nitrification Potential and Nitrifier Communities," *Front. Soil Sci.*, vol. 2, 2022, Accessed: Mar. 02, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fsoil.2022.838497.

[20] C. Soman, D. F. Li, M. M. Wander, and A. D. Kent, "Long-term fertilizer and crop-rotation treatments differentially affect soil bacterial community structure," *Plant Soil*, vol. 413, no. 1–2, pp. 145–159, Apr. 2017, doi: 10.1007/s11104-016-3083-y.

[21] R. T. Odell, *The Morrow plots : a century of learning.* Urbana, Ill.: Agricultural Experiment Station, College of Agriculture, University of Illinois at Urbana-Champaign, 1982. Accessed: Jul. 15, 2019. [Online]. Available: https://www.ideals.illinois.edu/handle/2142/8640.

[22] E. Bakker, F. Hubert, M. M. Wander, and B. Lanson, "Soil Development under Continuous Agriculture at the Morrow Plots Experimental Fields from X-ray Diffraction Profile Modelling," *Soil Syst.*, vol. 2, no. 3, p. 46, Aug. 2018, doi: 10.3390/soilsystems2030046.

[23] B. Velde and T. Peck, "Clay mineral changes in the Morrow experimental plots, University of Illinois," *Clays Clay Miner.*, vol. 50, no. 3, pp. 364–370, Jun. 2002, doi: 10.1346/000986002760833738.

[24] Y. Zhang, "100 Years of Pb Deposition and Transport in Soils in Champaign, Illinois, U.S.A.," *Water. Air. Soil Pollut.*, vol. 146, no. 1, pp. 197–210, Jun. 2003, doi: 10.1023/A:1023957226204.

[25] K. M. Wolf *et al.*, "The century experiment: the first twenty years of UC Davis' Mediterranean agroecological experiment," *Ecology*, vol. 99, no. 2, pp. 503–503, 2018, doi: 10.1002/ecy.2105.

[26] Board of Trustees, "Board of Trustees Transactions, 9th Report." Jun. 07, 1877.

[27] Board of Trustees, "Board of Trustees Transactions, 14th Report, Agricultural Experiment Station Report." 1888.

[28] Board of Trustees, "'Board of Trustees Minutes –1878-1880,' Board of Trustees Biennial Reports." 1880.

[29] "The Morrow Plots : a National Historic Landmark. Special Publications No. 16," Urbana, Ill. : University of Illinois College of Agriculture, Illinois, text Special publication (University of Illinois at Urbana-Champaign. College of Agriculture); no. 16, 1969. [Online]. Available: http://hdl.handle.net/2142/32745.

[30] Board of Trustees, "Board of Trustees Transactions, 1st Report, Laws Concerning the Industrial University, Laws of Congress, p. 1-3; 1st Report." May 07, 1867.

[31] B. E. Seely, "Engineering and the Land-Grant Tradition at the University of Illinois, 1868–1950," in *Science as service : establishing and reformulating American land-grant universities, 1865-1930*, M. Alan, Ed. Tuscaloosa: University of Alabama Press, 2015.

[32] "Morrow Plots Notebook, 1876-1913, 1967." Accessed: Mar. 02, 2023. [Online]. Available: https://digital.library.illinois.edu/items/b9a74f70-51c5-0138-7202-02d0d7bfd6e4-9.

[33] J. A. Kelly, S. L. Farrell, L. G. Hendrickson, J. Luby, and K. L. Mastel, "A Critical Literature Review of Historic Scientific Analog Data: Uses, Successes, and Challenges," *Data Sci. J.*, vol. 21, no. 1, Art. no. 1, Jul. 2022, doi: 10.5334/dsj-2022-014.

[34] D. Noonan and T. Chute, "Data Curation and the University Archives," *Am. Arch.*, vol. 77, no. 1, pp. 201–240, Jun. 2014, doi: 10.17723/aarc.77.1.m49r46526847g587.

[35] H. Wickham, "Tidy Data," *J. Stat. Softw.*, vol. 59, no. 10, pp. 1–23, Aug. 2014, doi: 10.18637/jss.v059.i10.

[36] S. Caldrone, "Morrow-Plots-Public GitHub Repository." Dec. 07, 2022. Accessed: Mar. 02, 2023. [Online]. Available: https://github.com/SandiCal/morrow-plots.

[37] C. Fallaw *et al.*, "Overly Honest Data Repository Development," *The Code4Lib Journal*, no. 34, Oct. 2016, Accessed: Oct. 25, 2021. [Online]. Available: https://journal.code4lib.org/articles/11980.

[38] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci. Data*, vol. 3, no. 1, Art. no. 1, Mar. 2016, doi: 10.1038/sdata.2016.18.

Digital Preservation and Accessibility of Archives in Oman:

Current Status and Future Directions

Ahmed Maher Khafaga Shehata
Sultan Qaboos University
Oman
a.shehata@squ.edu.om
https://orcid.org/0000-0002-5447-5867

Abderrazak Mkadmi
Sultan Qaboos University
Oman
a.mkadmi1@squ.edu.om
https://orcid.org/0000-0002-5621-2235

Abstract – Digital preservation significance is widely recognized and imperative for all institutions. It is a pressing concern for archival specialists who acknowledge its relevance and necessity in contemporary information management practices.

The current paper aims to examine the present state of digital preservation and Accessibility of institutional archives in Oman and to ascertain the challenges and prospects in this domain. This research is conducted through a comprehensive analysis of the extant literature and semi-structured interviews with administrative personnel employed in ~~both~~ public ~~and private~~ institutions in Oman. Semi-structured interviews were conducted with ten professionals in governmental organizations, who were chosen based on their expertise and experience in the field, using purposive sampling.

The findings showed that interviewees in Omani institutions value digital preservation for many reasons, such as compliance with Omani law on archives, improved organization and security, and ease of document access. However, all institutions have no unified application strategy for digital preservation. The findings revealed many challenges, including technological obsolescence, security risks, big data management, and human resource challenges.

Keywords – Digital preservation, Oman, Institutional archives, Information management, Accessibility

Conference Topics – DIGITAL ACCESSIBILITY, INCLUSION, AND DIVERSITY.

I. Introduction

Digital preservation and accessibility are increasingly important issues in the digital age, as the volume of digital content continues to grow and the reliance on digital technologies increases. In Oman, the adoption of digital technologies in a range of sectors, including education, business, and cultural heritage, has brought both opportunities and challenges for preserving and accessing digital information and materials.

Recent research has highlighted the importance of digital preservation in Oman for ensuring the long-term accessibility and usability of digital content (Aboraya et al., 2021). However, the country faces many challenges in this area, including limited technical infrastructure and expertise and a lack of standardization and interoperability among digital systems (Mehta & Hemmy, 2021). These challenges can make it difficult to preserve and access digital content over time and can hinder the ability of Omani organizations to take full advantage of the benefits of digital technologies.

Digital accessibility is also important in Oman, enabling individuals to participate fully in the digital world (Lucchi, 2013). However, a lack of awareness and understanding of the needs of people, as well as

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

a lack of accessibility standards and guidelines, can make it difficult to access and use digital content and services (Elnaggar, 2008; Hadidi & Al Khateeb, 2015; Kulkarni, 2019).

This conference paper aims to explore the current status of digital preservation and accessibility of archives in Oman and identify the challenges and opportunities in this area, through a review of the existing literature and semi-structured interviews targeting administrative staff working in public and private Omani institutions.

II. LITERATURE

A. Digital Preservation in Oman

Research on digital preservation in Oman has identified several challenges and opportunities in this area. Studies found that limited technical infrastructure and expertise and a lack of standardization and interoperability among digital systems are among the key challenges facing digital preservation in Oman. These challenges can make it difficult to preserve and access digital content over time and hinder Omani organizations' ability to take full advantage of the benefits of digital technologies (Al Hinai, 2016; Al Mughairi; Mehta & Hemmy, 2021).

Regarding opportunities, Al Hinai (2016) and Aboraya et al. (2021) identified the potential for digital preservation to support the digitization of cultural and heritage collections in Oman and improve the efficiency and effectiveness of government and business operations. However, they also emphasized the need for stronger policies and regulations to support digital preservation in the country and for greater investment in technical infrastructure and expertise.

In addition, digital preservation poses, according to (Mkadmi, 2021), several organizational, technical, legal, normative and strategic challenges. Technological obsolescence remains one of the main problems of information preservation and durability. And it is in this sense that archivists are called upon today and more than ever to find the necessary strategies for cooperation with other information professionals, namely computer scientists, lawyers, data analysts, auditors, etc., to emphasize both the preservation and the accessibility of the documents. Also at the normative level, the challenges are enormous, given the explosion on the one hand of standards in this field related to description, preservation, and accessibility and, on the other hand, the variety of documents to be preserved, which requires each a different strategy, while also putting in mind that other standards in other areas are also involved in this process, namely those of quality, computer security, human rights, etc.

Internationally, digital preservation has been recognized as an important issue for preserving and accessing digital content over time (Lee et al., 2002). Research has identified a range of challenges, including the need to ensure the authenticity and integrity of digital materials, the need to migrate digital content to new formats and technologies as they become obsolete, and the need to manage and store digital content in a way that ensures its long-term accessibility(Galyani Moghaddam, 2010; Gaur & Tripathi, 2012). Several approaches have been proposed to address these challenges, including using digital preservation frameworks and standards, developing digital preservation policies and strategies, and establishing digital preservation repositories and infrastructure (Becker et al., 2009; Masenya & Ngulube, 2020).

B. Digital Accessibility in Oman

Research on digital accessibility in Oman has highlighted the importance of this issue for enabling individuals with disabilities to participate fully in the digital (Abanumy et al., 2005; Al Sulaimani & Ozuem, 2022). However, the literature has also identified several challenges that can make it difficult for people with disabilities to access and use digital content and services in Oman. Studies found that a lack of awareness and understanding of the needs of people with disabilities, as well as a lack of accessibility standards and guidelines, are among the key barriers to digital accessibility in the country (Abanumy et al., 2005; Al Sulaimani & Ozuem, 2022).

Internationally, digital accessibility has been recognized as an important issue for ensuring that all individuals can access and use digital content and services (Jaeger & Xie, 2009; Valtolina & Fratus, 2022). Research has identified a range of approaches for promoting digital accessibility, including the development of accessibility standards and guidelines, the design of inclusive digital products and services, and the use of assistive technologies to support the needs of people with disabilities

(Kulkarni, 2019). To support digital accessibility, many countries have implemented laws and regulations requiring digital content and services to be accessible to people with disabilities (Lazar et al., 2015; Oliveira et al., 2020).

Despite these efforts, digital accessibility remains a challenge in many countries, including Oman. Research has identified a range of barriers to digital accessibility in Oman and Arab countries, including a lack of awareness and understanding of the importance of accessibility, a lack of trained professionals and experts in the field, and a lack of resources and infrastructure to support accessibility efforts (Mkadmi, 2021). Additionally, the rapid pace of technological change can make it difficult to keep up with the latest accessibility standards and practices. The lack of interoperability between different technologies can create additional barriers for people with disabilities (Kulkarni, 2019; Lewthwaite & James, 2020).

To address these challenges, it is important to adopt a holistic and inclusive approach to digital accessibility in Oman. This could involve a range of strategies, including the development of accessibility standards and guidelines, the design of inclusive digital products and services, the use of assistive technologies, and the implementation of laws and regulations to support accessibility efforts. By taking these steps, we can work towards ensuring that all Oman individuals can fully participate in the digital world.

2. The study significance

Digital preservation and accessibility importance has been widely acknowledged in Oman and internationally. In Arab countries, digital preservation has created numerous opportunities and challenges explored in previous research (Abubaker et al., 2015; Awamleh & Hamad, 2022). However, some gaps in the existing literature still need to be addressed to achieve more inclusive and equitable outcomes.

One area that requires further investigation is the impact of digital preservation and accessibility in Oman. There is a need for more in-depth studies on how digital technologies can benefit people to access archives and digital content. Understanding the challenges and opportunities of digital accessibility and preservation can help identify effective strategies for promoting access to information rights. Additionally, more research is needed on how digital technologies can promote cultural inclusivity and diversity in Oman.

3. Aims

- Highlight the importance of digital preservation and accessibility in Omani institutions.
- Identify the challenges and opportunities of digital preservation of documents and ways to access them in Omani institutions.
- Explore the potential future directions for digital preservation and accessibility in Oman, including emerging technologies and best practices that could help to ensure the long-term accessibility of digital content.

4. Methodology

To explore the current status of digital preservation and accessibility in Oman, we conducted semi-structured interviews with ten experts from ten Omani organizations. Our sample consisted of government r professionals, who were chosen based on their expertise and experience in the field using purposive sampling.

Table 1: Sample Details

Participant	Organization	Sector	Gender
1	Amman Airports	Government	Male
2	Oman Air	Government	Male
3	The General Secretariat of Tender Board	Government	Male
4	Ministry of Social Development	Government	Female

5	Anonymized	Government	Male
6	Environment Agency	Government	Female
7	Ministry of Justice and Legal Affairs	Government	Male
8	Omani Board of Medical Specialties	Government	Male
9	Ministry of Endowments and Religious Affairs	Government	Male
10	Namaa Holding	Government	Male

A. Data Collection and Analysis

The interviews were conducted in person by a research assistant specializing in digital Archives and lasted approximately 45 minutes each. The semi-structured interviews followed a set of guiding questions but also allowed for flexibility and exploration of additional topics that emerged during the interview. We recorded and transcribed the interviews for analysis. The interviews were in the Arabic language and then were translated into English.

We used thematic analysis to analyze the data from interviews, in which themes and patterns were identified and coded across the interview transcripts and grouped. This allowed us to explore the perspectives and experiences of our participants and gain insights into the current status of digital preservation and accessibility in Oman.

B. Limitations

One limitation of our study is the small sample size, which may not represent the variance of perspectives and experiences of all experts in digital preservation in Oman. Additionally, our interviews were conducted with professionals working in specific sectors and organizations, and therefore, the findings may not be generalizable to other sectors or organizations. However, our study provides valuable insights into the current digital preservation and accessibility practices in Oman and offers a starting point for further research.

5. Analysis

A. Importance of digital preservation

Regarding identifying the importance of digital preservation in Omani institutions, most respondents were well aware of what digital preservation is and its importance in business management, document control and retrieval when necessary. This is because most of the respondents have academic training in document management. and archives and they consider digital preservation as part of their job duties, which have absolute priority, in line with the speed and flow of documents in digital form. The respondent, A2, considered that "digital preservation helps in speedy retrieval of information in a timely manner, centralization of preservation, and provision of spaces occupied by paper documents. It also enables rapid sharing of information and greater confidentiality of information." In the same context, most respondents expressed their strong awareness of the importance of digital preservation as the best way to bypass spatial and temporal requirements (A5), as it enables the institution to facilitate access to documents at all times and from all places. It also allows more than one employee to access the same document, which avoids frequent copying of files, thus reducing operational costs (A4). However, one of the respondents (A6) considers that digital memorization is still somewhat recent in institutions and has started with the beginning of using e-mail (Outlook) as an official means of communication, which provides, in addition to messaging, other functions related to the calendar, task management, contacts, note-taking, and journal logging. This last respondent stated, "Before establishing archives departments in institutions, there was no digital preservation. The beginning of circulation of documents in digital form was since the institution adopted e-mail (Outlook) as an approved means of correspondence within the institution. The

management and preservation of digital records is still new to us."

In addition, the interviewees highlighted the importance of digital preservation in organizing and securing documents, especially since it was based on the archival tools prepared by the National Records and Archives Authority (NRAA), which is particularly represented in the document classification system and the Records retention schedules. Thus, (A8) considered that the process of "preserving digital files in the institution with organized and systematic applications linked to the classification system and schedule of retention periods of the NRAA contributes significantly to the rationalization and governance of document organization and retrieval processes." In this context, (A9) also expressed that digital preservation does not include preserving documents from loss but should also focus on protecting media and containers from the factors affecting them.

B. Digital preservation process

When we were asked how the digital preservation process takes place, whether it is carried out at the level of the institution or outside it or by adopting cloud computing, the answers were mixed. Most of the respondents mentioned that the archiving process takes place mostly inside the institutions, and sometimes it takes place in addition to that in cloud computing applications and other times in the servers of the ministry or institutions specialized in digital preservation outside the institution. A1, A2, A3, A5, and A9 mentioned that digital preservation is done on internal servers under the direct responsibility of the organization. In contrast, A2 and A3 mentioned that some documents are saved in cloud computing and internal archiving, citing Microsoft applications in this regard. On the other hand, A4 mentioned that the new direction is to work on finding a system that matches the standards of the NRAA, as he stated that "the archives department is trying, in its meetings with officials in the ministry, to persuade those concerned to implement a new system that is more compatible with the standards of the NRAA." In the same context, A8 also stated that it will "work on an electronic document management program in future, and it will be compatible with the standards of the NRAA." A6 also confirmed that the "Woussoul[1]" system prepared by the NRAA would soon be approved, which depends on the preservation policy on national cloud computing that will be concentrated within the Sultanate (in the Information Technology Authority in the Knowledge Oasis). However, A10 considers that digital archiving is still in its infancy, and there is no clear strategy for this purpose except for e-mail archiving.

C. Digital preservation strategy, tools and standards used

We believe the digital preservation process must be subject to a clear technical and administrative strategy. There must be an advanced technical structure that is subject to standards and specifications that guarantee the governance of documents and the speed of their retrieval while preserving their confidentiality, comprehensiveness and authenticity throughout the preservation period in the institution, as well as when implementing their final fate either by deporting them to the NRAA or by destroying them in accordance with the conditions and standards related to destruction. Therefore, our first question in this section is related to the existence of strategies and evidence on which digital preservation is based. The answers seemed to be distinct as well. Where more than half of the respondents (A4, A5, A7, A8, and A9) expressed the absence of a clear strategy and policy with regard to saving and securing data, and if there are some rules, then the information technology department alone follows up on that, without referring to the departments of documents management. In this regard, A4 stated, "There is no specific policy for keeping these files for the long term. With regard to risk management, data preservation and Backup, it is the responsibility of the information systems departments. The documents management department does not have into the process of signing the system contract. We have no idea about these operations and system updates". At the same time, the rest of the respondents (A1, A2, A3 and A6) confirmed that the digital preservation process is

[1] Woussoul (which means "access "in English) is the name of the "EDRMS" system which is developed by NRAA to generalize it thereafter to all the institutions concerned by the Omani law of archives.

subject to many procedures issued in the form of circulars and regulations by the relevant ministry and also by the Information Technology Authority and the Electronic Defense Center, where A3 stated that "it is following several procedures in the preservation process with regard to digital preservation, the circulars, policies, standards and regulations issued by the Ministry of Transport, Communications and Information Technology and the Electronic Defense Center are followed. For example, at the level of digital preservation on cloud computing, reliance is placed on the "cloud computing policy first" and the cloud computing governance framework".

As for the technological tools used in the digital preservation process, they are summarized, according to the respondents, in computers and databases within the electronic correspondence system, scanners, electronic sharing files, as well as photocopiers and servers for data preservation.

Specifications or standards of the digital preservation process was our third question in this section, by which we mean standards related to quality, metadata, structure, file formats, digitization, encryption, and standards related to access, including for people with special needs, etc. The answers to this question also ranged between positive and negative, as each of A1, A2, 3, and A5 confirmed that the archiving process is subject to standards related to metadata and file formats as well as to data encryption and that most of these standards were partially included in procedures guides developed by NRAA such as The National Guide to electronic documents and records management as well as by the Ministry of Transport, Communications and Information Technology. However, most of the respondents, including those who declared the existence of standards, did not provide details about the extent of the application of these standards or their contents. This is because only information technology employees deal with these issues and ensure their implementation.

D. Challenges of Digital preservation

It was found that digital preservation brings forth many challenges. One of the foremost challenges is technological obsolescence, which leads to electronic systems and information resources becoming outdated over time, thus making it difficult to use and fully leverage their potential. As expressed by Interviewee A7, "All electronic systems and information resources are subject to obsolescence, and their use becomes limited over time due to changes in programs and modern devices in the contemporary era, leading to difficulties in their utilization and optimal exploitation." Therefore, keeping up with modern changes, programs, and technologies is essential to overcome this challenge. In addition, digital preservation depends on the system developer rather than the institution, making renewing contracts critical in ensuring its maintenance over time.

Another significant challenge is security, as there is a risk of hacking, stealing, or erasing data, necessitating adherence to current security policies issued by relevant security authorities. Interviewee A7 emphasized the importance of "modern policies in the security of electronic documents and information issued by the security authorities such as the National Authority for cyber defense" and highlighted the Information Security Department's responsibility to ensure data safety. Other challenges include difficulty processing the increasing volume of data and the high flow of information, requiring expertise in information technology, high storage capacity requirements, and risk management for maintaining servers. Updates to the systems may also be unsuitable and not aligned with the nature of the data stored in old systems (A1).

Institutional and organizational challenges are apparent in the process of digital preservation. These include the presence of contracts with external companies, the lack of policies and guidelines to support the integration of documentation specialists with those in charge of digital preservation, and a shortage of specialized expertise in managing digital preservation, as Interviewee A3 pointed out, absence of policies that support specialists in integrating with those charged with digital preservation. Also, the resignation of expertise specialized in managing digital preservation." In addition, financial documents cannot be managed in digital form, which can lead to difficulties in understanding the context of the document for staff, as Interviewee A5 highlighted. Furthermore, the complexity of document management systems and the changing file formats

and sizes present further challenges, necessitating regular hardware updates. Access to the system should be limited to the organization to avoid the risks associated with using systems outside the organization's scope.

Human resource challenges, such as job instability, the failure to manage systems and the ambiguity of digital preservation standards, were also identified through the interviews (Interviewees A7 and A9, respectively). Overcoming these multifaceted challenges will require careful consideration of technical, security, institutional, and human factors. Ensuring effective and long-term preservation of digital documents will depend on addressing these challenges.

E. Opportunities of digital preservation

The interviews shed light on the potential benefits of digital preservation and the challenges it presents. According to the interviewees, digital preservation offers several opportunities, such as improved administrative memory, streamlined document sharing, and collaboration on the same file. The possibility of monitoring and auditing by many users was also highlighted (A2), as well as the potential to exploit all the advantages of digital technologies in processing, describing, and organizing data. Furthermore, digital preservation was essential for linking administrative institutions, facilitating services provision, and developing electronic and smart governments (A1).

According to Interviewee A3, digital preservation provides opportunities to support financial and administrative oversight inside and outside the institution, enhance project management, and improve spending efficiency. The presence of digital data and documents enables the acquisition of reliable decisions and reports and facilitates monitoring and following up on government projects. The interviewees also mentioned other advantages of digital preservation, such as lower space storage costs and easier data updates (A3).

In addition, the interviewees highlighted the benefits of easy and quick access to documents and files, which supports government efforts in improving services and enables institutional participation in administrative work. The accuracy, ease, and speed of file retrieval were also emphasized (A5 and A6).

Overall, the interviews suggest that digital preservation offers a range of potential benefits regarding data management, project management, and government services. Such benefits can be realized by overcoming the challenges posed by digital preservation, including technological obsolescence, security risks, big data management, high storage requirements, and human resource challenges. Addressing these challenges will be critical in ensuring an effective and long-term preservation of digital documents.

F. Future directions for digital preservation and accessibility in Oman

The data from the interviews indicate that the presence of a strategy for digital preservation in the organizations varied among the participants. While some organizations had a specific plan and strategy for digital preservation, others did not have a strategy in place.

Participants A1 and A2 confirmed that their organizations have plans to acquire a system for managing documents and electronic documents, which will be compatible with the standards of the NRAA. Participant A4 mentioned the general trend in the Sultanate to contract with the "Woussoul" program licensed by the Omani government, which is an integrated system for all institutions in handling digital files. Participants A5 and A6 acknowledged that the government's approach, through the launch of the government document management system (Woussoul), will lead to strong digital preservation in the future, which requires the development of special strategies to keep up with technological developments. Participant A8 also mentioned that their organization has a digital transformation team, indicating that the organization is taking steps towards digital preservation.

On the other hand, some participants noted the lack of a strategy for digital preservation in their organizations. Participants A3 and A7 confirmed that there is currently no strategy for digital preservation, but work is underway to implement the access system adopted by the NRAA. Participant A9 mentioned that the organization is developing legal regulations to facilitate the creation of a guide for managing digital content. Finally, participant A10 confirmed that their organization has a strategy for the future of digital preservation.

G. *Digital preservation specialists*

The responses to the question about designation of those responsible for digital preservation indicate that different organizations have varying approaches to this aspect. In some organizations, there is a designated person or department that oversees digital preservation, while in others, the responsibility is distributed among different departments or employees.

Institutions like A1, A2, A6, and A10 have designated individuals or departments responsible for digital preservation. A1, for example, mentioned that the organization has a director for the digital preservation Centre, and the IT officials are responsible for the electronic system. Similarly, A2 has a digital department that manages data preservation. Organization A6 has specialists in document management, and A10 has specialists in electronic preservation.

On the other hand, organizations such as A3, A4, and A8 do not have dedicated individuals responsible for digital preservation. In these cases, this task is distributed among different departments or employees. For instance, in A3, most employees of the IT department are responsible for digital preservation, while in 8a, all electronic systems are the responsibility of the IT department.

Overall, it is important for institutions to have dedicated individuals or departments responsible for digital preservation. Having such individuals or departments would ensure that the preservation process is given due attention and that the process is carried out efficiently and effectively.

H. *Adoption of new technologies*

From the interviews, it is evident that not all the organizations use cloud computing. Some of the organizations save their data on company servers, which means they do not use cloud computing (A1 and A4). In contrast, other organizations use cloud computing and have specified guarantees for the same within their service provision agreement (A3 and A8). In the case of organizations such as in A5 and A9 organizations, it is not clear if they use cloud computing or not.

Regarding the use of modern technologies related to big data and blockchain, most participants expressed some interest in adopting them. For instance, A2, A3, A6, and A8 expressed an interest in adopting modern technologies related to big data. However, some participants, such as A1 and A5, expressed no interest in adopting these modern technologies.

It is also essential to note that some participants, such as A4 and A6, do not clearly understand whether they use cloud computing. Those participants also do not clearly understand modern technologies related to big data and blockchain.

In conclusion, the use of cloud computing and adopting modern technologies related to big data and blockchain vary among organizations. Additionally, some organizations lack a clear understanding of whether they use cloud computing or not, and they also lack a clear understanding of modern technologies related to big data and blockchain. However, most organizations expressed a level of interest in adopting these modern technologies.

I. *Data Protection and Rights*

Based on the interviews, it appears that the participants hold varying opinions and knowledge regarding data protection and the right to be forgotten in the Sultanate of Oman. The data reveals that a Personal Data Protection Law exists in the Sultanate, which mandates entities to comply with its regulations (A1). However, some participants are unaware of the law or have insufficient knowledge about it (A4, A7). Others rely on the directives and policies issued by the competent authorities (A3), while some participants believe that a guide or similar regulation is unnecessary (A5, A9, A10).

One participant stresses the need for laws restricting loopholes in accessing personal data in the country (A6). Another participant points to the Royal Decree that pertains to the establishment of the NRAA, which specifies which documents can be viewed and when they can be viewed (A8). As for the existence of a strategy for the right to be forgotten, the interviews indicate that no such strategy exists in the Sultanate (A1, A2, A3, A5, A6, A7, A8, A9, A10). However, some participants noted that employee files might contain sensitive information subject to specific retention periods (A1, A8).

6. Discussion

It emerges from the results of our study that digital preservation is of interest to all the interviewees, especially who work in archives departments in public institutions. The majority are aware of the importance of this preservation and put forward at least three main reasons: first, the obligation to preserve all the administrative information stipulated by the Omani law on archives, and as all the institutions where our interviewees' work are public institutions and are directly affected by this law. Then, the participants consider that digital preservation guarantees the organization and security of documents, particularly by respecting the preservation and classification guides established by the NRAA. The third reason is the ease of finding and sharing documents: availability, immediacy and simultaneous access. All these reasons are consistent with what was presented in the study by (Aboraya et al., 2021) on the long-term accessibility and usability of digital content.

However, despite the importance given by archival specialists to digital preservation, the latter does not yet have a unified application strategy for all institutions. The results of our study show that storage is done both in internal servers and in private companies specializing in electronic archiving and by using cloud computing applications (including Microsoft). This prompted the NRAA to develop the "Wossoul" system with national cloud computing located inside the country, which all public institutions bet on to solve these problems of secure storage. This shows the awareness at the level of high authorities of the importance of digital preservation in an approach to the digital governance of institutions and the state.

Moreover, as this approach is only in its infancy, our questions on the tools, strategies and standards used in digital preservation have brought mixed results. Interviewees noted their optimism towards "Wossoul" system as a unified system with well-thought-out features and built-in standards related to metadata, encryption, file structure and format, and data security. These different standards are already mentioned in the various guides and recommendations of the NRAA and the Ministry of Transport, Communications and Information Technology.

The interviews shed light on the challenges Omani organizations face when attempting to preserve their digital documents, including technological obsolescence, security risks, big data management, high storage requirements, institutional and organizational challenges, and human resource challenges. Similar challenges were highlighted in the literature in many countries, including training and security risks (Kay Rinehart et al., 2014; Kirchhoff, 2008).

One of the main challenges identified is technological obsolescence, which leads to the rapid out datedness of electronic systems and information resources. This can make it challenging to use and fully leverage the potential of these resources, requiring organizations to keep up with modern changes, programs, and technologies to overcome this challenge. In addition, digital preservation depends on the system's developer, making renewing contracts critical in ensuring its maintenance over time (Conway, 2010).

Another significant challenge is security, as there is a risk of hacking, stealing, or erasing data, necessitating adherence to current security policies issued by relevant security authorities. This requires institutions to have recent policies to secure electronic documents and information issued by the security authorities, such as the National Authority for cyber defense.

Institutional and organizational challenges were also identified, including the lack of policies and guidelines to support the integration of documentation specialists with those in charge of digital preservation and a shortage of specialized expertise in managing digital preservation. Financial documents cannot be managed in digital form, which can lead to difficulties in understanding the document's context for staff. The complexity of document management systems and the changing file formats and sizes also present further challenges, necessitating regular hardware updates.

Human resource challenges, such as job instability of human cadres in information technology and the failure to manage systems and explain their details to specialists in the field of documents, were also identified through the interviews. Overcoming these multifaceted challenges will require careful consideration of technical, security, institutional, and human factors.

Despite these challenges, studies revealed many benefits, such as the storage of documents, ease of management, and providing powerful search options (Baro, 2016; Mannheimer & Cote, 2017). Similarly, our interviews shed light on the potential benefits of digital preservation, such as improved administrative memory, streamlined document sharing, and the ability to collaborate on the same file. Digital preservation was deemed essential for linking administrative institutions, facilitating services provision, and developing electronic and smart governments. The presence of digital data and documents enables the acquisition of reliable decisions and reports and facilitates monitoring and following up on government projects.

The literature reveals that emerging AI, blockchain, open data and internet of things technologies are considered among the top future directions of digital preservation (Adu et al., 2016; Hassan et al., 2019; Mannheimer & Cote, 2017). Future directions for digital preservation in Oman included developing specific plans and strategies for digital preservation, adopting modern systems for managing documents and electronic documents that are compatible with the standards of the NRAA, and using integrated systems for handling digital files. Organizations also need to develop special strategies to keep up with technological developments, and legal regulations must be developed to facilitate the creation of a guide for digital preservation.

7. Conclusion

The importance of digital preservation is not to be demonstrated; it is imposed today in all the institutions subject to our study, it is present and constitutes one of the most urgent concerns of the various archival specialists. Nevertheless, reflection at the national level must be done to put in place a national strategy for preservation and digital archiving in its broadest sense affecting both sectors: public and private. Our study shows that this work of strategic reflection and development of standards is already underway as part of the electronic government project. It remains to convince a certain timidity, on the one hand, to involve archivists with computer scientists in the various strategies and applications of digital preservation and, on the other hand, to take into consideration all the standards related to documents and which are in addition of those who deal with Records management, particularly those who are linked to the quality and security of both information and systems.

The challenges of digital preservation identified through interviews conducted in Oman include technological obsolescence, security risks, big data management, high storage requirements, and human resource challenges. These challenges call for careful consideration of technical, security, institutional, and human factors to ensure digital documents' effective and long-term preservation. On the other hand, the potential benefits of digital preservation include improved administrative memory, streamlined document sharing, easy and quick access to documents and files, lower space storage costs, improved spending efficiency, and enhanced project management. The data from the interviews revealed that the presence of a strategy for digital preservation in the organizations varied among the participants, with some having a specific plan and strategy for digital preservation while others did not have a strategy in place. Overall, there is a need for institutions in Oman to develop and implement strategies for digital preservation to realize the potential benefits and overcome the challenges posed by digital preservation.

Finally, a good archiving strategy, in our opinion, ensures the integrity of documents and accessibility in the short, medium and long term and participates in particular in preserving personal data and the privacy of individuals.

8. REFERENCES

Abanumy, A., Al-Badi, A., & Mayhew, P. (2005). e-Government Website accessibility: in-depth evaluation of Saudi Arabia and Oman. The Electronic Journal of e-government, 3(3), 99-106.

Aboraya, W., Shemy, N., Said, S., Alkalbani, M., Shehata, N., & Abdelhady, B. (2021). Investigating the necessity of having digital repositories in postbasic education in Oman. International Journal of Internet Education, 20(2), 13-24.

Abubaker, H., Salah, K., Al-Muhairi, H., & Bentiba, A. (2015). Digital Arabic content: challenges and opportunities. 2015 International Conference on Information and Communication Technology Research (ICTRC),

Adu, K. K., Dube, L., & Adjei, E. (2016). Digital preservation: the conduit through which open data, electronic government and the right to information are implemented. Library Hi Tech.

Al Hinai, A. S. (2016). Archives and its Role in Preserving the Nation Memory: Legal and Scientific Use of the Records and the Role of National Records and Archives Authority in Oman as a Model. Atlanti, 26(2), 197-208.

Al Mughairi, M. M. ARCHIVES AND THE DIGITAL AGE. International Institute for Archival Science of Trieste and Maribor, 38.

Al Sulaimani, A. H. A., & Ozuem, W. (2022). Understanding the role of transparency, participation, and collaboration for achieving open digital government goals in Oman. Transforming Government: People, Process and Policy(ahead-of-print).

Awamleh, M. A., & Hamad, F. (2022). Digital preservation of information sources at academic libraries in Jordan: an employee's perspective. Library Management.

Baro, E. E. (2016). Digital preservation practices in university libraries: A survey of institutional repositories in Nigeria. Preservation, Digital Technology & Culture, 45(3), 134-144.

Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., & Hofman, H. (2009). Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. International journal on digital libraries, 10(4), 133-157.

Conway, P. (2010). Preservation in the age of Google: Digitization, digital preservation, and dilemmas. The Library Quarterly, 80(1), 61-79.

Elnaggar, A. (2008). Towards gender equal access to ICT. Information Technology for Development, 14(4), 280-293.

Galyani Moghaddam, G. (2010). Preserving digital resources: issues and concerns from a view of librarians. Collection Building, 29(2), 65-69. https://doi.org/10.1108/01604951011040152

Gaur, R. C., & Tripathi, M. (2012). Digital Preservation of Electronic Resources. DESIDOC Journal of Library & Information Technology, 32(4).

Hadidi, M. S., & Al Khateeb, J. M. (2015). Special education in Arab countries: Current challenges. International Journal of Disability, Development and Education, 62(5), 518-530.

Hassan, M. U., Rehmani, M. H., & Chen, J. (2019). Privacy preservation in blockchain based IoT systems: Integration issues, prospects, challenges, and future research directions. Future Generation Computer Systems, 97, 512-529.

Jaeger, P. T., & Xie, B. (2009). Developing online community accessibility guidelines for persons with disabilities and older adults. Journal of Disability Policy Studies, 20(1), 55-63.

Kay Rinehart, A., Prud'homme, P.-A., & Reid Huot, A. (2014). Overwhelmed to action: digital preservation challenges at the under-resourced institution. OCLC Systems & Services, 30(1), 28-42.

Kirchhoff, A. J. (2008). Digital preservation: challenges and implementation. Learned Publishing, 21(4), 285-294.

Kulkarni, M. (2019). Digital accessibility: Challenges and opportunities. IIMB Management Review, 31(1), 91-98.

Lazar, J., Goldstein, D., & Taylor, A. (2015). Ensuring digital accessibility through process and policy. Morgan kaufmann.

Lee, K.-H., Slattery, O., Lu, R., Tang, X., & McCrary, V. (2002). The state of the art and practice in digital preservation. Journal of research of the National institute of standards and technology, 107(1), 93.

Lewthwaite, S., & James, A. (2020). Accessible at last?: what do new European digital accessibility laws mean for disabled people in the UK? Disability & Society, 35(8), 1360-1365.

Lucchi, N. (2013). The Role of Internet Access in Enabling Individual's Rights and Freedoms. European University Institute-RSCAS Working Paper(2013/47).

Mannheimer, S., & Cote, C. (2017). Cultivate, assess, advocate, implement, and sustain: A five-point plan for successful digital preservation collaborations. Digital Library Perspectives.

Masenya, T. M., & Ngulube, P. (2020). Factors that influence digital preservation sustainability in academic libraries in South Africa. South African Journal of Libraries and Information Science, 86(1), 52-63.

Mehta, S. R., & Hemmy, K. (2021). Digital Humanities in Oman: Logistics, Challenges and Opportunities. Knowledge Cultures, 9(1), 56-74.

Mkadmi, A. (2021). Archives in the Digital Age: Preservation and the Right to be Forgotten. John Wiley & Sons.

Oliveira, A. C., da Silva, L. F., Eler, M. M., & Freire, A. P. (2020). Do Brazilian Federal Agencies Specify Accessibility Requirements for the Development of their Mobile Apps? XVI Brazilian Symposium on Information Systems,

Valtolina, S., & Fratus, D. (2022). Local Government Websites Accessibility: Evaluation and Finding from Italy. Digital Government: Research and Practice, 3(3), 1-16.

RUNNING UP THAT HILL

Making Digital Preservation Skills Accessible with Novice to Know-How

Sharon McMeekin
Digital Preservation Coalition
Scotland
sharon.mcmeekin@dpconline.org
0000-0002-1842-611X

Melinda Haunton
The National Archives (UK)
United Kingdom
Melinda.Haunton@nationalarchives.gov.uk
0000-0002-4885-6313

Abstract – Over the last three years, The National Archives (UK) and the Digital Preservation Coalition have collaborated on the development of a growing body of training content under the banner of the "Novice to Know-How" learning pathway. The content has proved to be incredibly popular, opening the door to digital preservation training for individuals and organizations that previously had not been able or willing to engage with the topic. This paper will examine the projects motivations, outputs, impact, and future plans.

Keywords – Training, Workforce Development, Skills, Collaboration

Conference Topics – We're all in this together; From theory to practice; Digital accessibility, inclusion, and diversity

I. INTRODUCTION

At iPres 2019 in Amsterdam, the authors of this paper both presented within the same session, each reporting on their organization's approaches to helping build digital skills capacity.

Melinda discussed The National Archives (UK)'s sector leadership role and a recent survey carried out in partnership with Jisc. The survey aimed to assess levels of skill and confidence across a range of digital activities. She detailed how the results of the survey were driving the development of a soon to be published strategy for building capacity within the UK Archives Sector [1],

Sharon's paper focused on the challenges of developing digital preservation skills for individuals and organizations, and existing and future work from the Digital Preservation Coalition (DPC) to help. She also suggested a number of areas where potential resources and/or collaboration within the digital preservation community could help more effectively meet the challenges faced [2].

Little did they know that they would soon be collaborating on a major training project that would begin to address the issues raised. The project in question would produce the popular "Novice to Know-How: Digital Preservation Skills for Beginners" (N2KH) learning pathway.

In this paper we will set out the motivations for the project, describe the project's execution, and discuss feedback and impact, before looking to the future.

II. WHY AN ONLINE COURSE LIKE NOVICE TO KNOW-HOW?

The context for Novice to Know-How is drawn from Plugged In, Powered Up (PIPU), the digital capacity building strategy published by The National Archives (UK) in 2019. The need for the strategy was based on sector research into skills, capacity, confidence and resourcing of archive services and archives professionals to carry out activity across core digital delivery areas: preservation, access, and engagement. A survey of over 300 archives workers in 2019 [3] was foundational to understanding in-depth requirements within The National Archives (UK)'s area of responsibility; a range of further in-person events and statistical analyses expanded and provided evidence to underpin the strategy.

Focusing on digital preservation, the 2019 strategy included key findings such as:

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

- Case for wide, general upskilling to benefit archives across the spectrum where TNA needs to deliver: "Unfortunately a gap has developed (and appears to be widening) between institutions leading on digital preservation and the remainder of the sector." (p18)
- Case for training to support wide staff understanding of concepts and tools rather than simply purchasing systems: "Software is not a substitute for knowledge and archives undertaking a procurement exercise instead of developing their in-house expertise in digital preservation risk simply spending money on tools they do not really understand." (p20)
- Case for urgency and risk to collections from limitations of skills: "48% of respondents reported they could not generate a checksum of a digital file, 49% could not perform file format analysis and 55% could not extract and publish metadata from a digital file. In each case, roughly another 25% of respondents reported that they 'had some knowledge/skills' in the specified area. This amounts to a very worrying proportion of the sample of the profession being unable to carry out critical preservation functions on digital records. This is so deeply concerning because these findings amount to an admission that the nature of contemporary collections is such that today many archive professionals can no longer care effectively for the material they hold." (p20-21)
- Case for hands-on, practical walk throughs and detailed tool support rather than focusing on needs at a policy level – there was already high quality training available, which clearly made the case for needing to approach digital preservation, but archives staff who had attended such training frequently did not report an increase in their confidence or skills to handle digital material in practice: "Training must also be the right sort of training. It certainly cannot be purely theoretical…, digital preservation is a craft skill and must be learned in practice as well as theory." (p21)
- Case for easy access: the survey did not specifically seek out information about online training, but the issues of time, opportunity and (to a lesser extent) cost all emerged, demonstrating a lack of commitment from parent organizations to support archivists in gaining skills that they had identified as critical (p24). An online training course would not address all these issues, but it could reduce some of the barriers and introduce flexibility of scheduling.

This collected evidence overwhelmingly supported developing an online training course, alongside other skills development opportunities such as peer mentoring, an in-person "Archive School" in which trainees learned from digital preservation specialists at The National Archives (UK), and the creation of reference materials [4].

Internal funding was secured from The National Archives (UK) to seek external support to create the online learning pathway. The invitation to tender emphasized a number of points drawn from the research and which became core to delivering the learning program. Participants were expected to have low initial skills and confidence and should build these over time during their learning. The emphasis was on practical and hands-on learning, around a variety of tools, which trainees could then implement in their own workplaces. The role of collaborative working with colleagues in IT was emphasized. There was also a strong desire to connect the learning pathway with best practice already in existence. The tender referenced DigCurV and the Digital Preservation Handbook as possible mapping tools to achieve the right level of skills and coverage of the subject.

III. Developing Novice to Know-How

In a coincidence that would later feel like kismet, at the same time The National Archives (UK) was evaluating the digital skills of the UK Archives Sector and making plans for development, the DPC was reevaluating their approach to training and development.

The DPC's "Getting Started…" and "Making Progress with Digital Preservation" courses had long been a cornerstone of the organization's training provision. The training courses offered learners a broad introduction to the range of activities required to establish a digital preservation program, from developing policy to designing workflows. The courses were each held three times a year at venues

across the UK and Ireland and always received positive feedback from attendees.

While the courses were popular with attendees, their potential impact was limited. They were only available to a maximum of 180 learners a year within a limited geographical area. When considered in light of the internationalization of the DPC, this approach to training raised serious concerns in relation to accessibility and ongoing sustainability.

Developing self-directed online training surrogates for the "Getting Started…" and "Making Progress…" courses was identified as the best option moving forward as this would allow more learners access to the materials while being time zone agnostic. There were, however, significant barriers to making this plan a reality. The DPC would need to procure a Learning Management System (LMS), staff would have to develop the skills required to author training content suitable for delivery online, and additional capacity would be required to facilitate the time-consuming process of creating that online content. As a small, non-profit organization, the DPC lacked the resources to move forward with these plans.

The announcement of the "Invitation to Tender" from The National Archives (UK) was, therefore, both timely and exciting. If the project tender could be secured it would not only provide the opportunity for the DPC to make a move into the world of online training, but also to do so in partnership with a long-time ally and friend.

The DPC's tender proposal was submitted in November 2019, and notification of its success was received shortly afterwards. Within a project deadline of 31st March 2020, work was quickly initiated.

DPC staff joined colleagues at The National Archives (UK) in Kew, London, over the 11th and 12th of December 2019, while the rest of the UK was focused on a general election, to develop learning objectives and a course structure for N2KH. This process included consideration of the outcomes of the Digital Skills Survey, along with a wide-ranging analysis of digital skill requirements as indicated by several digital preservation good practice resources. The resources examined included the DigCurV Framework [5], The NDSA Levels of Preservation [6], the CoreTrustSeal [7], and the DPC's own Rapid Assessment Model [8].

The structure developed would become known as the N2KH "Learning Pathway", which included, in the first instance, six courses and 24 modules. The structure was designed specifically for those with little or no digital preservation knowledge, aiming to provide them with the skills needed to put basic workflows in place at their organization. Once the course structure had been determined the process of content creation began.

It had been agreed by the two organizations that the training content should be delivered in a variety of formats that would help engage different types of learners. To facilitate this and to gain the general skills required for creation of content ready for online delivery, DPC staff undertook research on and training in good practice for online training. This included learning around the range of products contained within Articulate's 360 software suite, a market-leading product for authoring online training content.

To ensure that the content developed was suitably clear, engaging, and authoritative, a robust approach to drafting, review, and update was undertaken. Each module was researched and drafted by its author before evaluation by at least two reviewers drawn from The National Archives (UK) and DPC staff. Edits were then made before a final review was undertaken.

Volunteers were also recruited for a pilot of a selection of the training materials to assess the suitability of content and its delivery. A target of 30 pilot participants was originally set, but 109 expressions of interest were received. Ultimately, 58 pilot participants were invited to evaluate the test materials. A survey and focus group were used to capture feedback from the participants and changes to content were made in response to their comments.

During this time a procurement exercise was also undertaken to identify a suitable LMS. This included drafting of requirements, identification of potential systems from the large number of options available in the LMS marketplace and testing of three systems which best met the identified requirements. SAP's LITMOS was ultimately chosen for delivery of N2KH.

Version 1.0 of the N2KH learning pathway content was delivered to The National Archives (UK) on time for the 31st of March deadline. Version 1.0 contained modules covering the following topics:

1. Introduction to Digital Preservation
2. Files, Files Formats, and Bitstream Preservation
3. Using DROID
4. Select and Transfer Digital Content
5. Ingesting Digital Content
6. Preserving Digital Content

They aimed to provide a balance of the theory behind digital preservation work and practical, actionable advice for those who were new to the topic. There was also an emphasis on free or low-cost solutions that would be accessible to those with few available resources. Content was delivered in a range of formats including video, text, interactive elements, click-through tool demos of DROID, and short quizzes.

A beta launch was offered in April, with early access provided to pilot volunteers (both those who participated and those who were not selected). The learning pathway was officially launched at an online event on 4th of May 2020, with the first monthly cohort of learners beginning the course on the 1st of that month. N2KH is offered for free to all learners, with priority places available each month to learners from the UK Archives Sector and from DPC Members.

IV. N2KH Reception and Learner Feedback

Given the results of The National Archives and Jisc survey, it was expected that N2KH would be popular, but the level of enthusiasm for the learning pathway was a surprise to all of those involved in its development. Places in the first monthly cohort of 140 learners sold out in less than one day. The DPC immediately increased the number of places available with additional financial support received from The National Archives (UK).

The number of registrations received can be partially attributed to timing of the learning pathway's release, just as the world was entering the first COVID-19 lockdown, but three years on it still remains incredibly popular. As of 9th March 2023, 2734 learners have now completed N2KH. And while the largest group of learners have been UK-based, as befits a course developed with the UK Archives Sector in mind, the N2KH learning pathway has been undertaken by learners from 62 countries across Africa, Asia, Australasia, Europe, North America, and South America.

Completion rates are also high, with an average of over 65% over the lifetime of the learning pathway. This is much higher than rates observed for online courses generally. One study from the Open University found that the median completion rate for online courses of 12.6%, with the highest of those included in the study being 52.1% [9].

Extensive feedback has been gathered to ensure that the learning pathway is meeting the needs of learners and to help guide future updates and development. To capture this feedback, learners are invited to complete a survey which includes questions on:

- Their level of digital preservation knowledge
- How long they took to complete the learning pathway
- If the content was appropriate for a novice/beginner level
- What content types they preferred
- If they found the knowledge check quizzes useful
- How they found navigating between resources
- What was done well
- What was missing
- If they encountered any errors (e.g., spelling, and broken links)
- Any other comments or feedback

To date, feedback has been very positive. Over the lifetime of the project the following has been observed:

- Around 95% of respondents have "strongly agreed" or "agreed" that content of the modules was at an appropriate level
- Around 90% of respondents have "strongly agreed" or "agreed" that there is a good variety of content types
- Respondents particularly enjoyed the quiz and tool demo elements
- More practical exercises and case studies were the most wished for additional content

There have also been strong themes within the textual answers provided around:

- Appreciation for the course structure and how content is split into easily digestible portions.
- Praise for the clarity and simplicity of the explanations within the learning pathway,

breaking down the complex issues of digital preservation into content that was accessible.

- A level of apprehension felt by many learners about digital preservation before undertaking the learning pathway, and how they felt a confidence now to begin facing the challenges.
- Offering thanks for the availability of course, noting that they had struggled to either gain access to other training dues to barriers such as time or funding, or with the higher levels of presumed existing knowledge of other courses.

The following quotes are taken from the feedback and are representative of the comments received:

> *I felt the course was extremely well structured. Key concepts were explained from the ground up, allowing us to build up a good knowledge base from strong foundations.*

> *It was the first thing on DP that I felt I could understand and that addressed the practicalities more than the theory.*

> *I enjoyed the whole course and I found it went back to basics to clearly explain fundamental points and build up, which filled in some knowledge gaps for me. It has made me more confident that digital preservation is something that I could now do rather than just aspire to.*

> *I really felt that this was a practical course – I was able to go away from each module and think about it in the context of my organisation and start to experiment with small steps using the knowledge I had learnt through that day's module.*

Fewer than 1% of learners have replied to the feedback survey with negative comments. These have mostly been from those that felt the level of the course was still too high for them.

V. Further Content Development and Content Sharing

In light of the positive reception of N2KH 1.0, The National Archives (UK) has subsequently open invitations to tender for additional rounds of content development. The DPC's tender proposals have been successful on each of these occasions and the two organizations have now worked in partnership on a number of N2KH projects that further enriched the training content.

Later in 2020 a small project known as N2KH 1.1 was undertaken to add further practical elements to the course, as requested in the feedback. This took the form of four new tool demos, focusing on free or low-cost tools for fixity checking, moving/copying content, and characterization, and a fifth module on using the command line. The project also included formatting of these modules, and existing modules on using DROID, for delivery via the "Digital Preservation Handbook" [10] as well as through the learning pathway. This was done so that the content would be available to learners after completion of N2KH, whilst also making them available to the wider digital preservation community.

A second major N2KH project was commissioned in early 2021, with the aim of adding a new course on "Providing Access to Preserved Digital Content", rounding out the learning pathway's coverage of the digital lifecycle. Access had been out of scope of the original N2KH learning pathway due to time constraints and the limited number of good practice publications available. But the publication of the DLF's "Levels of Born-Digital Access" [11] and other resources now made the creation of content a possibility. The new course was added to the existing learning pathway in April 2021, and is also offered separately for those who had previously completed the original N2KH.

Most recently, The National Archives (UK) and the DPC have started a new N2KH project (3.0), this time focusing on delivering training specifically on the topic email preservation. This learning pathway aims to be a progression on the original learning pathway, assuming a solid foundation of general digital preservation knowledge from learners. Again, the content will look to provide a solid theoretical foundation whilst also offering practical skills and advice that can be put into action. In response to the feedback received from learners, real-world case studies are also being included within the learning content. It is expected that the learning pathway will be open to learners in July 2023.

In addition to the new training content that has been and is being developed, work has been undertaken to share the training content with other organizations for deployment within their own learning management systems. All of the N2KH modules have been created in the interoperable learning content packaging standard SCORM 1.2. A guide to the use of the learning content and its structure has been created and all content is available under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license[1]. As of March 2023, nine organizations have uploaded N2KH their own LMS. These organizations include a number of universities and national collecting organizations, with the content being used to contribute to both teaching and internal staff development.

Support has also been provided on an ad hoc basis for groups wishing to engage with N2KH as a single cohort. This has ranged from simply offering administrative support to organize access to N2KH at the same time for the group, rather than requiring individual registrations, through to DPC staff providing additional training workshops on topics complementary to the main N2KH learning pathway. Additional topics covered have included policy development, risk management, continuous improvement, and advocacy. Informal feedback has suggested this has been a positive experience for those taking part. The benefits mentioned have included:

- Additional motivation and support gained from sharing the learning experience with colleagues
- Improved clarity and outcomes when working on digital preservation thanks to a shared foundation of knowledge

Due to the positive responses received to the complementary sessions, the DPC will soon be trialing a series of "Novice to Know-How Plus" sessions for their Australasia and Pacific members.

VI. "Plugged In, Powered Up" Review

In 2022, Simon Wilson, an experienced archives consultant, was employed to conduct a review of Plugged In, Powered Up to gauge the impact of the range of activities provided across the capacity building strategy. This did not directly seek feedback on N2KH as standalone, as the training course was more widely circulated beyond TNA's leadership activity, but it was a key element of the activities to be assessed. The key to success of PIPU would be to see progress in those areas of digital skills and confidence across the UK archives sector. As with the 2019 Jisc/TNA survey, another sector-wide survey was explored and understood in more depth through focus groups. 172 responses were received. Almost 3 in 4 had completed N2KH, the largest recognition and participation level of all the PIPU outputs to date.

The survey outcomes show a major change as a result of 3 years of digital capacity building (Table 1). From 34% agreeing "I have sufficient digital expertise to deliver my role" in 2019, by 2022 this had increased to 43%. A remarkably consistent 9% improvement was also seen in responses to the contrasting "My colleagues have insufficient digital expertise" (agreement reduced from 49% to 40%).

Other key outcomes include:

- 12% improvement in services where digital archives are stored in multiple, geographically distinct locations (41% to 53%)
- 11% improvement in regular fixity checking (17% to 28%)
- 12% improvement in having a digital strategy in place (35% to 47%)
- 12% improvement in clear responsibility for digital preservation (33% to 45%)
- 19% improvement in offering access to digital records of some kind (37% to 56%)

	2022	2019

[1] https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

	No knowledge	Have the knowledge	No knowledge	Have the knowledge
Generating a checksum of a digital file	17% ↓	44% ↑	47%	26%
Performing file format analysis	22% ↓	38% ↑	48%	24%
Managing permissions for digital files	21% ↓	19% ↑	48%	17%
Extracting metadata from born-digital files	33% ↓	16% ↑	55%	15%
Redacting a document for web publication	46% ↓	13% ↓	54%	18%

Table 1: Survey results in relation to specific skills

(Note: figures are approximate as slightly different scales were used across the two surveys)

There is evidently much room to continue to improve, but this is a significant shift in a short period of time – which was also, of course, a highly disrupted period for all working in archives services, when the exigencies of dealing with the pandemic took precedence at times, and where some archive services were entirely closed for months if not longer, with employees on furlough.

Some respondents were keen to see Novice to Know-How content more accessible outside the online learning management system, as they wanted reference access. This is in fact already feasible: tool demos are embedded into the Digital Preservation Handbook, and the entire course or selected elements downloaded as reference materials, for participants, or as a whole for those who have not taken the course. This emphasizes how valued and flexible the content developed through N2KH is, although it also underlines the need to continue to communicate the alternate ways in and ongoing accessibility of the content in different contexts. There was spontaneous appetite for an "archivist to digital archivist learning pathway", suggesting that the skills development and confidence-building approach of N2KH had real resonance for participants.

One might get depressed about the levels of real confidence in practical tasks but Wilson states "It is interesting to note that the confidence level for the practical tasks…is higher than it is for the advocacy and broader tasks." – in general there is very low confidence, and it is not especially focused on the practical at it was in 2019.

The 2019 survey showed a high demand for online training, which was not visible in the 2022 survey, presumably because this had been developed and delivered to such a high proportion of respondents.

13 respondents to the 2022 survey identified their main barrier in delivering more digital preservation as having the confidence to follow-on from Novice to Know-How training – this underlines that future developments need to retain a focus on confidence, and not go too far too fast if they remain targeted at mass audiences. Many more respondents wanted "more of the same" from TNA's next steps.

Further anecdotal evidence of the positive impact of N2KH has been observed in the applications for UK Archive Service Accreditation since 2020. The majority of applications have cited N2KH as a means of improving digital preservation skills across archive services, bringing skills up to a shared level and informing colleagues to the point where institutional conversations can move on.

Less positively, there was still a sense of frustration and lack of confidence in digital preservation skills for certain respondents. Some commented that N2KH itself assumed too much and "a more basic introduction to digital archives is needed to introduce the key concepts". Given that

N2KH starts from a very basic understanding that digital records require management, this was dismaying. This may underpin the analytical observation that the previous gap between best and worst preservation performers is widening. The weakest services may require a different approach. It may also be linked to issues with general information technology competencies and related confidence within the sector.

Additionally, only 19% of respondents say they have a complete digital asset register (41% say no/don't know). This may in part be an artefact of using different terminology in N2KH but is a worrying baseline.

VII. Future Plans

Continuing to deliver an online training offer which goes from basics to more expert is a clear priority for both The National Archives (UK) and the DPC.

At present we have focused on novice to intermediate work, with the forthcoming email preservation learning pathway offering the first in a potential series of content looking at challenges of particular formats.

It may be that such courses to expert level are worthwhile, but The National Archives (UK) does not yet have a clear steer from the archives sector as to what would be most valuable, so they will continue to explore options and monitor sector needs. At present, all development of N2KH has required competitive tendering, and project funding, often on contracts with tight deadlines. This has created opportunities which would otherwise have been impossible but also management issues and time pressures. It may be that an alternative collaboration model could work better in the long term.

The National Archives (UK) hope to continue to support networking and skills sharing, through peer mentoring and opportunities to share staff expertise with the sector. They are also aware that capacity for development time in their target audiences is limited and are considering whether there is a role for a national body in modelling and advocating time for training across the country as well as a clear role in advocating for the significance of the work

There is also the question of how to serve those who are not sufficiently confident even for N2KH. This may involve additional support for particular activities, such as getting to a Digital Asset Register baseline.

In line with their ongoing program of translation of other key resources, the DPC has considered the possibility of translation of the N2KH to languages other than English. Unfortunately, there are additional barriers to this process for online learning content. The time, skill, and knowledge required to carry out the translations are not the only resource requirements, for online learning there are also significant resource requirements and complications resulting from the need to specially format the content in the correct format for delivery through an LMS. Although it is unlikely that progress with translations will be possible in the short-term, the DPC will continue seek possible opportunities.

The organization is also working towards the delivery of online training development outside of the N2KH collection. Courses are already in development on the topics of "Continuous Improvement" and "Risk Management for Digital Preservation" and additional courses around a variety of digital preservation topics are being considered. Without additional funding, the DPC will not be able to make this training free to all as with N2KH, but it will be considering different funding models that will ensure the content is financially accessible whilst meeting costs.

The continued sustainability and relevance of N2KH will remain a key priority for both organizations. A yearly review schedule has been included in forward planning to ensure the content continues to represent good practice in the ever-evolving field of digital preservation.

The learning pathway will also remain free for all learners, although the number of monthly places available may be reduced if additional funding to support access is not secured. Further promotion of the course to new audiences is also planned, as are efforts to increase awareness of the availability of the content within local LMSs.

VIII. Conclusion

The development of the Novice to Know-How learning pathway has been a rewarding endeavor for all those involved. It has reached a large, international audience, and has provided many practitioners with the grounding in digital

preservation practice they have previously been unable to access.

The National Archives (UK) and the DPC will continue to collaborate on supporting the sustainability of the learning pathway and will investigate the possibilities for further developments that will help broaden access to digital preservation knowledge and practice.

ACKNOWLEDGEMENTS

We would like to acknowledge the essential contributions colleagues at The National Archives (UK) and the DPC have made to the development of Novice to Know-How and its successes, from its initial conception through to the upcoming third anniversary of its launch.

In particular, this paper is presented in memory of Dr. Jo Pugh, The National Archives (UK), a devoted advocate for building digital skills within the UK Archives and Cultural Heritage Sectors, and without whom Novice to Know-How would not exist.

REFERENCES

[1] Haunton, Melinda, Pugh, Jo, Travers, James. "Building Network Capacity Among Memory Institutions: A Multi-strand Development Approach." Proceedings of the 16th International Conference on Digital Preservation, Amsterdam, September 2019. https://osf.io/w75cb/

[2] McMeekin, Sharon. "People Get Ready: Building Sustainability into Workforce Development." Proceedings of the 16th International Conference on Digital Preservation, Amsterdam, September 2019. https://osf.io/dtqe8/

[3] The National Archives (UK) and Jisc, "Archival Workforce Digital Skills Report", April 2019, https://www.contractsfinder.service.gov.uk/Notice/Attachment/ed93ed2f-e6b5-44d6-b65a-67593f2adc08

[4] The National Archives (UK), Digital Preservation Workflows, https://www.nationalarchives.gov.uk/archives-sector/projects-and-programmes/plugged-in-powered-up/digital-preservation-workflows/

[5] The DigCurV Competency Framework, https://digcurv.gla.ac.uk/skills.html

[6] National Digital Stewardship Alliance, NDSA Levels of Digital Preservation, https://ndsa.org/publications/levels-of-digital-preservation/

[7] The CoreTrustSeal, https://www.coretrustseal.org/

[8] Digital Preservation Coalition, Rapid Assessment Model, http://doi.org/10.7207/dpcram21-02

[9] Jordan, Katy, "Massive open online course completion rates revisited: Assessment, length and attrition", International Review of Research in Open and Distributed Learning, 16(3) pp. 341–358 http://oro.open.ac.uk/43566/

[10] Digital Preservation Coalition, "The Digital Preservation Handbook", 2nd Edition, https://www.dpconline.org/handbook

[11] DLF Born-Digital Access Working Group, "Levels of Born Digital Access", https://osf.io/r5f78/

PRESERVING ONLINE JOURNALISTIC CONTENT IN DISRUPTIVE TIMES

The case of collection.news

Lok Hei Lui
University of Toronto
Canada
kenlh.lui@mail.utoronto.ca
0000-0001-5077-1530

Abstract – Journalistic content is a crucial part of history, yet its longevity always remains uncertain without proper curation and preservation. This is true in particular when it comes to journalistic content under authoritarian regime contexts, where freedom of the press and information freedom are usually in vain. The article explores the case of collection.news, a community initiative that crawled, disseminated and hosted the journalistic content of Apple Daily, a pro-democracy media outlet that was forcibly shut down by the authority in Hong Kong. By discussing the key events and tools used by collection.news initiative, the three distinctive features of it, namely exigency, decentralization, and anonymity, are highlighted. Finally, suggestions to the digital preservation field for supporting these community initiatives in authoritarian regimes will be given.

Keywords – collection.news, archives-at-risk, authoritarianism, community archives, Hong Kong

Conference Topics – Digital accessibility, inclusion and diversity

I. INTRODUCTION

There is a common saying that "journalism is the first draft of history". News content is one of the important records that document the events happening around the world, and also has become an indispensable part of many people's lives. That said, freedom of the press is not guaranteed in many parts of the world, especially for people living in authoritarian regimes. Critical journalism platforms operating under such regimes are always being targeted by the authorities since true journalism, which involves exposing government wrongdoings, could be a threat to these regimes. When these media platforms are cracked down by the regime, the associated news content, if not well preserved by third parties, is usually vanished.

This article will discuss the case of collection.news. The project is a community initiative that preserved the web content of Apple, a now-defunct pro-democracy media platform based in Hong Kong. The research methods will be outlined in the next section. After that, the key events, tools used, and approaches adopted of the initiative will be illustrated and then the analysis follows. Lastly, a conclusion will be drawn and suggestions for the digital preservation field will be provided.

II. RESEARCH METHODS

The paper adopts a qualitative approach in this study by analyzing primary and secondary sources. These sources include forum posts, collection.news website and its GitHub repo documentation. Drawing upon the analysis, the author will further discuss the tools used, coordination and distinctive features of the preservation project, make analysis, and give suggestions.

III. BACKGROUND OF APPLE DAILY

Apple Daily was a prominent pro-democracy media outlet before its forced closure in June 2021. In 2019, Hong Kong experienced the largest-scale pro-democracy movement, the Anti-Extradition Bill Movement, in the territory. In response to the political unrest in Hong Kong, the Chinese government promulgated the controversial Hong Kong National Security Law (NSL) on July 1, 2020, the 22nd anniversary of the handover of Hong Kong's

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

sovereignty from the United Kingdom to the People's Republic of China.

Pro-democracy media platforms were deemed as one of the high-risk groups being targeted by the authority under NSL [1]. Two months after the NSL came into effect, Jimmy Lai, the founder of Apple Daily, was arrested by the National Security Department of the Hong Kong Police Force on suspicion of "collusion with foreign forces". On the same day, the police also searched the headquarters of Apple Daily. Despite the arrest, Apple Daily kept its business as usual afterward.

However, less than a year later, on June 17, 2021, the National Security Police arrested other management of Apple Daily and searched the headquarters again. The authorities also froze Apple Daily's assets, which eventually led to the media's cessation of operations. On June 23, the board of Apple Daily announced that the company would terminate its operations no later than June 26, and its digital platform would be shut down by midnight June 24.

IV. The Emergence of Collection.news

Following the forced shutdown of Apple Daily on June 24, 2021, a netizen "五大素球缺汁不可" (user id: #355204) created a thread on LIHKG forum, which is a Reddit-like forum based in Hong Kong, announcing that they had web-crawled over four hundred thousand articles from Apple Daily's website [2] (Fig. 1). In addition, the original poster expressed their wish to index the content afterward for web hosting and invited other forum users to contribute ideas on how to distribute and host the content. Some forum users suggested in the thread that there should be a frontend website for hosting the news article content with search functionality, while others proposed some distribution methods/platforms such as InterPlanetary File System (IPFS), BitTorrent, and GitLab for disseminating the news content data.

Figure 1 Snapshot of the inaugural thread discussing the crawling of Apple Daily web content on the LIHKG Forum

Two days later, the original poster created another thread [3] and included a GitHub repository link [4] to the initiative's documentation. On the GitHub repository, the author outlines the initiative's position and aims. Below is a translated version:

- The initiative primarily aims to back up the textual content of Apple Daily as I strongly believe in the power of words.
- The initiative aims to index the content and host an SEO-friendly website for people to search for old articles.
- Revealing the data is meant to promote brainstorming and encourage us to think about how we can utilize the data. The initiative does not intend to conceal the data. In fact, most people will not extensively browse the data after it has been backed up.
- Revealing the data can achieve the goal of decentralization. Even if someone who possesses the data gets into trouble later on, others can still continue.
- The initiative does not intend to crawl all of Apple Daily's content, such as images, videos, Instagram accounts, YouTube channels, Telegram channels, Facebook accounts, etc. I am aware that someone else is working on this.

The GitHub repository also provided a tutorial on how the end-users could download a copy of the media content through the Resilio Sync download tool in the forum post. The author explained the adoption of Resilio Sync: because of its decentralized, high-speed P2P sharing and flexibility in the modification of source files features.

V. COLLECTION.NEWS FRONTEND ACCESS AND ITS FUNCTIONALITY

Less than a month after the original post, on July 21, 2021, the same user created another thread on the LIHKG forum. They mentioned that after some effort throughout the weeks, they crawled more than 2.2 million articles from Apple Daily's website and hosted a website [1] for frontend access to news articles. The original poster also mentioned that the aim of the website is hoping an essential part of Hong Kong history would not be faded out because of the closure of Apple Daily.

Fig. 2 is the landing page of Apple Daily's content on collection.news website. By clicking on the boxes, users will be directed to the corresponding article. Akin to the layout of Apple daily's original website, the top grey bar lists different categories of articles. The date selection menu, represented by the middle black box (選擇日期), allows users to sort articles based on their publication dates.

In the top right-hand corner, users can access the website's search function. The indexing service is provided by Google (Fig. 3). This feature allows users to search articles by keywords. Mentioned in the FAQ section of collection.news website, using Google's indexing service is based on financial considerations and mitigating the risk of experiencing cyber attacks on the indexing server by hackers.

Figure 2 Screen capture displaying the landing page of Apple Daily's content on collection.news

Figure 3 Screen capture showcasing the search functionality on collection.news

VI. STRENGTHS AND LIMITATIONS OF COLLECTION.NEWS

The Internet Archive is another important platform for archiving Apple Daily's web content. However, the two platforms serve different functionalities, and have their pros and cons. The left and right sides are screenshots of the same article from the Internet Archive's Wayback Machine and collection.news respectively (Fig. 4). In comparison, the most significant advantage of collection.news is its ability to showcase attached photos, an essential

Figure 4 Screen capture comparing Wayback Machine and collection.news platforms for the same article

[1] https://collection.news/

component of online news articles, whereas Wayback Machine was unable to crawl the picture for this article and many other instances.

However, one major problem for the collection.news platform is that the content might not be up-to-date and may affect the data integrity of the content. As shown in the timestamp, Wayback Machine successfully crawled a more recent version of the webpage (2021.04.16 17:57), whereas the content hosted on collection.news was from an earlier version (21.04.16 02:00), which was 15 hours earlier. This discrepancy is most likely due to the limitation of web crawling from a legacy source before the complete shutdown of Apple Daily's web server.

Also, despite collection.news being a newly-built website for hosting the archived news content of Apple Daily with search functionality, it is unable to preserve the user interface and layout of Apple Daily's website, unlike Wayback Machine does.

VII. Analysis of the Preservation Process

Table 1 Summary of tools used by collection.news categorized by usage

Frontend Access	collection.news
Content Distribution	Resilio Sync, IPFS
Announcement/Coordination	LIHKG forum
Documentation	GitHub, GitLab

Table 1 summarizes the tools and platforms adopted by collection.news project. In contrast to traditional institutional approaches to implement preservation projects with long-term planned, structured and centralized features, the whole collection.news digital preservation project was an autonomous, decentralized and anonymous digital preservation movement initiated by passionate netizens. The three distinctive characteristics of this community-led project are exigency, decentralization and anonymity, respectively.

Exigency is one notable characteristic of this project. The Apple Daily web content was an archives-at-risk with only a small window of time to plan and execute the preservation process. From Apple Daily being searched by the National Security Police on June 17,2021, to the time that Apple Daily eventually ceased operation by midnight June 24, 2021, there was less than a week of time. This tight timeframe posed challenges to the preservation project facilitators, since they would have to work under intense pressure and grasp the golden period before the complete shutdown of service to crawl the data as much as they could. This urgency also meant that the preservation plan was likely to be incomplete and rough, potentially leading to critical data loss.

Decentralization is also another distinctive feature. Most digital preservation projects, due to financial, management and staffing considerations, are usually managed by GLAM (Galleries, Libraries, Archives, and Museums) institutions with a centralized operational approach, whereas the community-led collection.news project was operated in a decentralized way:

1. For preservation storage, there was no centralized data repository or platform for long-term preservation. Instead, the project publicly disseminated the news content data to end-users and relied on every single end-user for long-term preservation. This practice was entirely different from most institutional centralized approaches.
2. For data dissemination, the collection.news initiative made use of peer-to-peer protocol tools such as Resilio Sync and IPFS to disseminate the news content. The main advantage is the decentralized feature that could disseminate data with multiple users simultaneously while avoiding download speed bottlenecks. Another benefit of using peer-to-peer protocols is to prevent government internet censorship or denial-of-service attacks on a single hosting platform.
3. Adopting GitHub and GitLab as the platforms for documentation was also a decentralized approach. These open-source project platforms enable open collaboration and backup of content from every user without restrictions. This can ensure further access to the documentation. Also, similar to the case of Mainland China internet users, hosting documentation and organizing community archives on GitHub could be a way to circumvent Chinese government internet censorship [5].

Another distinctive characteristic of this project is its emphasis on anonymity. While most digital preservation projects were organized by identifiable institutions or organizations, collection.news

initiative was largely operated under the radar. The user name of the original poster's account on the LIHKG forum was a pseudonym. The GitHub repo was also owned by a brand new, designated account with no prior history. In addition, the initiative never publicly recruited volunteers nor openly organized crowdfunding campaign for funding. The organizational and operational details, such as funding, the number of facilitators and the decision-making model, remain concealed. This was, as mentioned in the FAQ on collection.news, intended to reduce the potential political risks.

VIII. Conclusion

This article introduces the case of collection.news, an autonomous and anonymized community initiative for preserving the online content of a Hong Kong-based newspaper platform, Apple Daily. The article then overviews the preservation process and approaches adopted by collection.news, by highlighting the key events and tools used. In the later part, this article points out three distinctive features of the whole community initiative compared to conventional digital preservation projects, namely exigency, decentralization, and anonymity. This case study should be helpful for readers to understand community-led digital preservation activism issues under authoritarian regime contexts.

In recent years, the world has been experiencing serious global democratic backsliding. With the expansion of authoritarianism, unfortunately, there might be a growing trend of more cases like the sudden collapse of Apple Daily. With reference to the distinctive features of collection.news discussed in the previous section, the digital preservation field could take specific actions to support these community initiatives:

1. Exigency: Authoritarian governments' crackdown on their targets is always unexpected. It is critical to plan ahead to collect and preserve the records and data before they vanish. While there is relatively sophisticated development for research data management cycles and digital preservation lifecycles, such as the DCC's Curation Lifecycle [6] and DPC's preservation Lifecycle [7], our field should consider developing standalone lifecycle frameworks for community digital preservation projects. These frameworks could help civil society actors, especially those in authoritarian regimes, to plan in advance and avoid abrupt crackdowns that leave little time for preservation efforts, just like in the case of collection.news.
2. Decentralization: Decentralization is an effective way to rapidly and widely disseminate censored data while mitigating political risks. However, without central management, the longevity and integrity of these digital assets remain uncontrollable and uncertain. To address this challenge, institutions from the free world could provide storage and techniques for parties to relocate their endangered digital materials. One example is Safe Havens for Archives at Risk Initiative [8], which is dedicated to providing support to organizations or individuals that need to deposit their records documenting human rights violations in reliable repositories.
3. Anonymity: Anonymity is crucial when it comes to conducting archiving initiatives in authoritarian regimes, as it ensures the safety of the initiative's facilitators. To support the facilitators of these community initiatives in circumventing state surveillance, more tutorials and technical assistance should be provided to teach them how to use encryption platforms and tools, such as Tor [9] and Session [10], for communication and operation without being detected by state authorities.

ACKNOWLEDGEMENTS

The author wishes to thank the guidance provided by Steve Marks, and, in particular, Jess Whyte from the University of Toronto Libraries along the way. Also, the author wants to salute the unsung heroes who took the initiative to preserve such precious records of Hong Kong's history.

REFERENCES

[1] Amnesty International, "Hong Kong: Targeting of pro-democracy newspaper is threat to press freedom," *Amnesty International*, Aug. 10, 2020. https://www.amnesty.org/en/latest/news/2020/08/hong-kong-targeting-of-pro-democracy-newspaper-is-threat-to-press-freedom/ (accessed Jun. 24, 2023).

[2] "[現場直播]我 backup 左蘋果四十萬篇文章 | LIHKG," *LIHKG 討論區*, Jun. 24, 2021. https://lihkg.com/thread/2588517/page/1 (accessed Mar. 10, 2023).
[3] "話說我之前 backup 左蘋果四十萬篇文章想大家幫手 download | LIHKG," *LIHKG 討論區*, Jun. 26, 2021. https://lihkg.com/thread/2591598/page/1 (accessed Mar. 10, 2023).
[4] appledailyarchive, "collection-news/appledaily-archive-directory: 蘋果日報文字備份目錄," *GitHub*, Nov. 29, 2021. https://github.com/collection-news/appledaily-archive-directory (accessed Mar. 10, 2023).
[5] A. Acker and L. Flamm, "COVID-19 Community Archives and the Platformization of Digital Cultural Memory," presented at the Hawaii International Conference on System Sciences, Jan. 2021. doi: 10.24251/HICSS.2021.312.
[6] Digital Curation Centre, "Curation Lifecycle Model," *Digital Curation Centre*. https://www.dcc.ac.uk/guidance/curation-lifecycle-model (accessed Mar. 10, 2023).
[7] Digital Preservation Coalition, "Preservation Lifecycle," *Digital Preservation Coalition*. https://www.dpconline.org/digipres/tags/preservation-lifecycle (accessed Mar. 10, 2023).
[8] Safe Havens for Archives at Risk Initiative, "Safe Havens for Archives at Risk Initiative," *Safe Havens for Archives at Risk Initiative*. https://safehavensforarchives.org/en/about-the-initiative/ (accessed Mar. 10, 2023).
[9] The Tor Project, "The Tor Project | Privacy & Freedom Online." https://torproject.org (accessed Mar. 10, 2023).
[10] Session, "Session | Send Messages, Not Metadata. | Private Messenger," *Session*. https://getsession.org/ (accessed Mar. 10, 2023).

Calculating the Carbon Footprint of Digital Preservation

A Case Study

Mikko Tiainen
CSC – IT Center for Science
Finland
mikko.tiainen@csc.fi
https://orcid.org/0009-0000-8513-6262

Juha Lehtonen
CSC – IT Center for Science
Finland
juha.lehtonen@csc.fi
https://orcid.org/0000-0002-9916-5731

Heikki Helin
CSC – IT Center for Science
Finland
heikki.helin@csc.fi
https://orcid.org/0000-0003-4002-8203

Johan Kylander
CSC – IT Center for Science
Finland
johan.kylander@csc.fi
https://orcid.org/0000-0002-8084-8233

Abstract – Environmental sustainability is becoming an important factor in digital preservation. We have calculated the carbon footprint of our Finnish national digital preservation services, which we provide for cultural heritage and research sectors. We concentrate on the carbon footprint of manufacturing hardware and shipping the equipment to data centers, and the carbon footprint of the hardware service life and employees related to the services. Using data provided to us by the hardware manufacturers and other sources, we show that the majority of the emissions come from manufacturing and shipping of hardware, whereas the emissions created during the service life has a smaller role. As a whole, the annual carbon footprint of the services is smaller than the annual carbon footprint of three average Finns.

Keywords – sustainability, carbon footprint, data centers, hardware manufacturing, hardware service life

Conference Topics – Sustainability: Real and Imagined

I. Introduction

Our national digital preservation repository, funded by the Ministry of Education and Culture of Finland, provides services for preserving cultural heritage and research data [1]. Our concept includes two national services: (1) The Digital Preservation Service for Cultural Heritage (in production since 2015) preserves digital assets from the cultural heritage sector, represented by archives, libraries and museums, and (2) The Digital Preservation Service for Research Data (in production since 2019) preserves data from the research sector, represented by universities and other research institutes. Given the diversity of the user needs, the digital assets to be preserved make up a very heterogeneous whole while simultaneously requiring various and flexible solutions. Both of these services together are in this paper referred to as Digital Preservation Services (DPS). The technical solution behind the services is common for both cultural heritage data and research data.

The carbon footprint of an IT-service can typically be modeled by breaking the service down to its separate components. The hardware has a lifecycle carbon footprint starting from manufacturing the raw materials, transportation of the hardware, production usage, and lastly the disposal of the hardware. On a data center level, data center power usage effectiveness (PUE) [2] is the driving factor together with hardware electricity usage when calculating the production usage carbon footprint. Enterprise level hardware vendors provide their own figures for the carbon footprint for their hardware.

In addition to the footprints mentioned above, the employee footprint includes emissions from

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

offices, traveling, heating, waste management and so on. The employee footprint consists of carbon emissions resulting from the daily work of administrating, developing, and managing the DPS.

We calculate the total carbon footprint of our DPS in this paper. These calculations apply only to our current configuration and thus cannot be applied in general to any other DPS. They might however provide some general guidelines and insights for others.

For calculating the carbon footprint, our services can be divided into hardware, data centers, network, administration work, development work, and supporting ICT-services. The carbon footprint of constructing the data centers is not within the scope of our calculations: DPS's should in general be geographically distributed to several data centers. The density of data storage is now on a level where only a few server racks are needed to hold several petabytes of data. Therefore, our DPS does not need its own data centers and we utilize only a minor part of the existing data centers. The data centers thus facilitate many other IT-services in addition to our DPS.

The paper is divided as follows: In Chapter 2 we describe the hardware of our DPS, in Chapters 3 and 4 we show the carbon footprint of manufacturing and shipping the hardware and of the actual service life, in Chapter 5 we bind these findings together with some observations, and in Chapter 6, we conclude the paper with future work.

II. Hardware

Our DPS platform utilizes three separate data centers for storing preserved copies in order to reduce geographical risks. The available capacity of the DPS is currently 3.6 peta bytes per copy. Currently, the platform consists of the following hardware:

- 13 x HPE Proliant DL360 Gen10 frontend and validation servers (ingest)
- 10 x HPE Apollo 4510 Gen10 storage servers
- 4 x HPE Apollo 4200 Gen9 tape library front end servers
- 2 x HPE Apollo 4200 Gen10 tape library front end servers
- 2 x IBM TS4300 tape library with 7 IBM full height LTO-8 tape drives and 336 LTO-8 tapes
- 1 x IBM TS4300 tape library with 7 IBM full height LTO-9 tape drives and 231 LTO-9 tapes
- For tape drives 15 % duty cycle is estimated.

Our DPS platform also includes a dark archive storage for mitigating worst case disasters related to online storage copies. The dark archive can be divided into three components when making calculations about its footprint: (1) Dark archive copy manufacturing, (2) Dark archive copy logistics, and (3) Dark archive copy storage.

We are not required to have dedicated resources for dark archive logistics and storage as they are shared with multiple other customers. Logistics are organized into monthly transports to the dark archive.

The volume of the dark archive is the same as our DPS platform. Currently the dark archive consists of LTO-8 tapes stored in Pelican 1450 transport cases. The total number of these cases is 24, and the total number of dark archive LTO-8 media is 336.

III. Manufacturing and Shipping

The carbon footprints of hardware manufacturing and shipping (more accurately: raw materials, manufacturing, shipping, and disposal at the end of the life cycle) have been reported to us by the manufacturers. The calculations from both of the manufacturers are based on the Product Attribute to Impact Algorithm (PAIA) [3] and represent the status of the products in 2022. From these given calculations, Table 1 summarizes the carbon footprint of our DPS platform for hardware manufacturing and shipping.

The hardware components used in our DPS platform for ingesting and preserving contents can be divided into different roles. We can calculate the carbon footprint of the DPS platform based on the following roles: ingest, spinning disk storage, magnetic tape storage, and the dark archive. Fig. 1 depicts the relative size of the carbon footprint from manufacturing and shipping for each hardware role.

The carbon dioxide emissions for the last mile of transportation need to be calculated separately because the distances from the manufacturer sites

to our data centers are different. The HPE servers are shipped to our data centers from within the EU while tape hardware is shipped from North America. The HPE servers are thus shipped into Finland by ground and sea transport whereas IBM tape hardware is transported via air. As an example, delivering a fully equipped IBM TS4300 tape library with seven full height LTO-8 tape drives to Finland has a logistics carbon footprint of 1449 kg CO_2ekv. In comparison, the logistics carbon footprint for ten HPE Apollo 4510 Gen10 servers is 675 kg CO_2ekv. These calculations are included in the sums in Table 1.

Table 1. The carbon footprints of manufacturing and shipping per unit.

Component	**Number of devices**	**Carbon footprint (kg CO_2ekv)**
HPE Proliant DL360 Gen10 (ingest)	13	14079
HPE Apollo 4510 Gen10 (spinning disk)	10	44890
HPE Apollo 4200 Gen9 (LTO-8, LTO-9)	4	9404
HPE Apollo 4200 Gen10 (LTO-8, LTO-9)	2	4704
IBM TS4300 tape library (LTO-8, LTO-9)	3	19116
IBM LTO-8 tape drives	7	665
IBM LTO-8 tape media (active, dark)	672	5020
IBM LTO-9 tape drives	7	665
IBM LTO-9 tape media	231	1726
Pelican 1450 transport case (dark)	24	210
Total DPS platform manufacturing CFP		**100479**

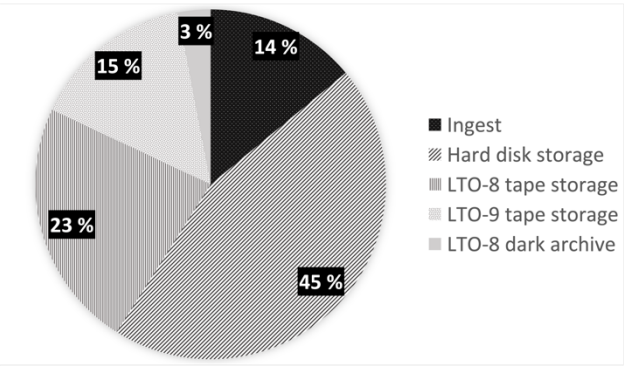


Figure 1. Carbon footprint division of manufacturing and shipping.

The calculation of the carbon footprint for the dark archive contains emissions resulting from manufacturing the LTO-8 media tapes and the Pelican 1450 transport cases. The exact carbon footprint of a case has not been provided to us, but we can estimate it by looking at the materials from which the case is constructed. A case weighs 2.5 kg and its raw material is polypropylene. Our figures are estimated from the carbon footprint of polypropylene pipe manufacturing [4] and they consist of producing polypropylene molecules and manufacturing the case. The total carbon footprint for manufacturing a Pelican 1450 case is estimated to be 8.4 kg CO_2ekv. This is an insignificant part of our whole carbon footprint.

IV. Service Life

The carbon footprint of the hardware service life depends on data center Power Usage Effectiveness (PUE). Currently, our services are located in three separate data centers with different PUE values: (1) Data center A with a PUE value of 1.66; (2) Data center B with a PUE value of 2; and (3) Data center C with a PUE value of 1.2. The PUE value defines the energy efficiency of the data center. For example, a PUE value of 1.2 means that the data center requires 20% energy on top of the real power usage of the DPS platform. It can for example be cooling or lighting. The electricity production for the data centers is done with Finnish hydropower where the corresponding carbon dioxide emission is 24 kg CO_2ekv / MWh. This figure is based on information found in the carbon footprint calculation tool created by the Finnish Environment Institute [5].

Table 2 depicts the carbon footprint for each hardware component of our DPS during its service life. The calculations include the PUE of the data center where the components are located. We assume in our calculations that servers with hard drives have a lifespan of five years while tape libraries and media have a lifespan of seven years. Fig. 2 shows the relative size of the carbon footprint of the service life for each role of the hardware: ingest, spinning disk storage and magnetic tape storage.

We have in close collaboration with our partner organizations (organizations that preserve their data in our DPS) defined common national preservation specifications, which in detail describe how digital

Table 2. Service life carbon footprint.

Component	Number of devices	Service life (years)	Data Center	Data Center PUE	Annual electricity (kWh)	Service life carbon footprint kg CO_2ekv
HPE Proliant DL360 Gen10	13	5	A	1.66	1358	3517
HPE Apollo 4510 Gen10	10	5	A	1.66	1209	2408
HPE Apollo 4200 Gen9	2	5	B	2	3320	1594
HPE Apollo 4200 Gen9	2	5	C	1.2	3320	956
HPE Apollo 4200 Gen10	1	5	B	2	2812	675
HPE Apollo 4200 Gen10	1	5	C	1.2	2812	405
IBM TS4300 tape library	2	7	B	2	5472	1838
IBM LTO-8 tape drives	7	7	B	2	2711	911
IBM LTO-8 tape media	672	7	B	2	0	0
IBM TS4300 tape library	1	7	C	1.2	2736	552
IBM LTO-9 tape drives	7	7	C	1.2	2711	547
IBM LTO-9 tape media	231	7	C	1.2	0	0
	Summary of usage time carbon footprint					**13402**

assets should be prepared before ingesting them to the preservation service. This includes for example requirements for metadata and file formats. We put a lot of effort into automated validation of the submission information packages and their assets during the ingest phase: This includes for example virus checks, full metadata validation, file format validation and verification of checksums. Our service also performs continuous monitoring of integrity by calculating and verifying checksums. For all these operations, to mention only a few, we use the GlusterFS distributed file system[1], MongoDB databases[2], Python programming language, and various 3rd party open source components. Our software stack as a whole uses 100% open source solutions.

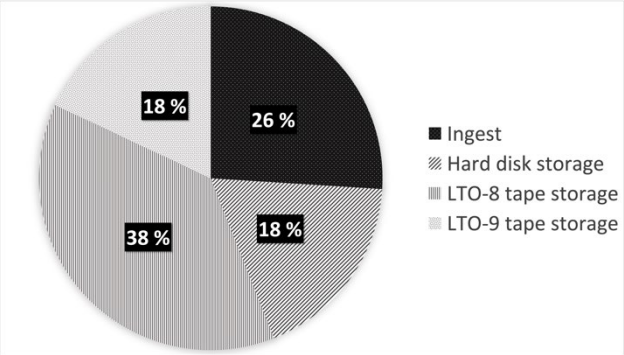


Figure 2. Service life carbon footprint division of hardware.

Our DPS have 17 experts working full time. The employee carbon footprint is calculated to have been 17.14 kg CO_2ekv in 2021, making our total annual carbon footprint for human resources in our services 292 kg CO_2ekv.

The carbon footprint of the dark archive is close to zero. We use one transport case per month, which makes the carbon footprint for the logistics around 4 kg CO_2ekv per year. Two years are needed to transfer all 3.6 peta bytes into the dark archive using LTO-8 tapes. The storage facility is located in a natural environment, shared with many other users, where external temperature and humidity control is not needed.[3]

[1] https://www.gluster.org/
[2] https://www.mongodb.com/

[3] The PUE value is therefore effectively 1.

V. OBSERVATIONS

Due to low carbon dioxide emissions of electricity production, the service life carbon footprint is only around 14% when compared to the manufacturing and shipping carbon footprint. This is shown in Fig. 3. This ratio is low even though we put a lot of computing resources into the validation of the submitted content during ingest. The majority of the carbon footprint is thus generated during manufacturing and shipping, and not during the actual service life of the hardware.

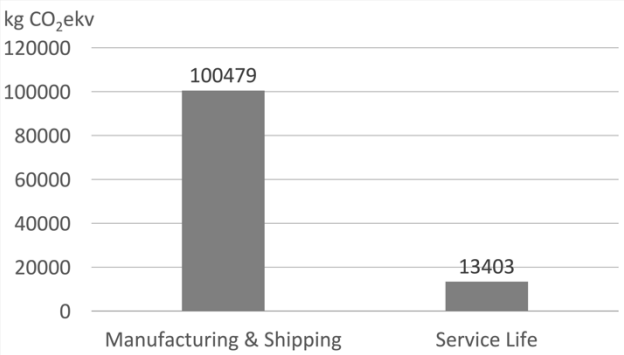


Figure 3. Manufacturing and shipping create a large carbon footprint compared to the service life.

When considering the storage areal density impact on the carbon footprint, the spinning disk areal density has the highest density and therefore its lifetime carbon footprint is not that far away from the footprint of tape environments. LTO-8, which has the lowest areal density, suffers from the fact that two modular tape libraries are needed to handle 3.6 peta bytes of storage.

Using electricity production with lower carbon dioxide emissions decreases the carbon footprint and reduces the impact that data center PUEs have on the total carbon footprint. Another major point of view that needs to be considered is however the total energy consumption during operation, regardless of the carbon footprint produced by it.

Table 2 shows the different life spans for the storage solutions. The annual carbon footprints for the different storage solutions with their differing life spans taken into account in the figures are shown in Table 3.

Table 3. The annual carbon footprint.

Component	**Annual carbon footprint kg CO_2ekv**
Ingest	3520
Spinning disk storage	9460
Magnetic tape storage (LTO-8)	4532
Magnetic tape storage (LTO-9)	3092
Dark Archive	273
Human resources	292
Total annual carbon footprint	**21169**

It can be noted that the dark archive with LTO-8 magnetic tapes has the lowest annual carbon footprint by far of all hardware components. Active tape environments suffer from tape servers that read and write the data, producing emissions in doing so.

The electricity production emissions play a role, if not a decisive one, in the total carbon footprint. Obviously, electricity production with low emissions should be prioritized.

As a collective result, our annual DPS carbon footprint is 21169 kg CO_2ekv. The Finnish Innovation Fund Sitra has calculated the average annual carbon footprint for a Finnish citizen in 2018, concluding that it is 10300 kg CO_2ekv [6]. The total carbon footprint of our DPS amounts to the carbon footprint of slightly less than three average Finnish citizens on an annual basis.

VI. FUTURE WORK

A few missing components from the calculation have been recognized: The results do not yet include carbon emissions of the optical network, data communication, or common support components for production and development. The carbon footprint relating to pre-ingest processing of digital content is also not within the scope of this paper.

The current calculations will become outdated when we increase the storage capacity or update the hardware. Carbon footprint calculations should be updated regularly whenever hardware infrastructure is changed or renewed.

Some possibilities to reduce carbon footprint are for example changing disk storage to other storage technology with a lower carbon footprint, favoring environment friendly technology and data centers, using emission free electricity, aiming for high areal

density in storage media, and increasing the service life of hardware components in use.

By the end of 2023 the ingest and spinning disk components will be transferred from our site in southern Finland to Northern Finland. The data center cooling in the new site is implemented with open air free cooling which leads into an excellent PUE of 1.05. This means an annual reduction of 435 kg CO_2ekv to our carbon dioxide emissions.

We have not utilized Green Coding [7], but the possibility to reduce carbon footprint through efficient processing is something to consider in the future.

A large work is ahead for IT-infrastructure manufacturers. They have to learn to minimize their products manufacturing carbon footprint. One component in this would be extensive recycling of product materials. A second major change which will have a significant impact is the green energy transformation for the production phase of IT-hardware. This transformation has just started in Europe and the future is promising regarding this shift.

We as consumers must start to require and prioritize more environment friendly infrastructure. Hopefully, the digital preservation community and IT experts are able to find ways to influence this in a positive way.

REFERENCES

[1] Finnish National Digital Preservation Services. https://digitalpreservation.fi/en/

[2] Wikipedia: *Power Usage Effectiveness*. https://en.wikipedia.org/wiki/Power_usage_effectiveness

[3] Massachusetts Institute of Technology, Material Systems Laboratory: *Product Attribute to Impact Algorithm - PAIA*. https://msl.mit.edu/projects/paia/main.html

[4] Jeroen Wassernaar, *Polypropylene Materials for Sewerage & Drainage Pipes with Reduced Energy and Carbon Footprints*, Journal of Materials Science and Engineering B6 (11-12), 2016.

[5] Finnish Environment Institute: *Y-HIILARI Hiilijalanjälki -työkalu* [carbon footprint tool]. https://www.syke.fi/fi-FI/Tutkimus__kehittaminen/Kulutus_ja_tuotanto/Laskurit/YHiilari

[6] The Finnish Innovation Fund Sitra: *Carbon footprint of the average Finn*. https://www.sitra.fi/en/articles/carbon-footprint-average-finn/

[7] Green Software Foundation: *Principles of Green Software Engineering*. https://learn.greensoftware.foundation/

THE VALORIZATION OF THE TUNISIAN RADIO ARCHIVE IN THE ARTIFICIAL INTELLIGENCE ERA

Sami Meddeb
Digital Cooperation Association Tunisia
Tunisia
s.meddeb@s2t.tn

Randi Cecchine
Independent Researcher
The Netherlands
Randi.Cecchine@gmail.com

Abstract – Radio Elyssa's sound archives face challenges in the post-revolution Tunisian context: preserving a large volume of digital content with limited resources while preserving Tunisia's cultural heritage. The Digital Cooperation Association Tunisia and Radio Elyssa are collaborating to test the potential of Artificial Intelligence tools to automate archival tasks. Before any automation, it is necessary to understand the regulatory procedures for digital archiving: identifying, classifying, migrating, and storing data according to best practices. Long-term preservation faces challenges due to limited resources and increasing digitization. The combination of human expertise and AI will enable Radio Elyssa to use AI responsibly to fulfill its mission.

Keywords – artificial intelligence, image, music, Audiovisual, chatbot.

Conference Topics – immersive information, We're All in this Together

I. INTRODUCTION

The period since the 2011 revolution in Tunisia has presented multiple challenges in the capacity to preserve and make accessible Tunisia's rich audiovisual cultural heritage. In response to the various challenges facing Tunisian institutions as well as the new potentials of digital preservation, the Digital Cooperation Association Tunisia ("DCAT") was formed in 2019 by members of the audiovisual archives community. DCAT supports digital archival efforts through partnerships, research, and trainings (such as Webinar Series "Digitisation and Restauration of Audio-visual Objects" organised by Landesarchiv Baden-Württemberg) aimed at developing archival expertise and building a collaborative community. This article describes a partnership between DCAT and a community radio station called Radio Elyssa that focuses on researching and testing the potential of artificial intelligence tools to support digital preservation strategies and audience engagement. Radio Elyssa was created in 2012 to present events and educational programs from the city of Gabes, with a focus on civil society and sustainable development activities of the municipality. The project tests how appropriate use of artificial intelligence tools can help to make archives more accessible, better organized, and preserved, thus highlighting and realizing their cultural value.

II. INTEGRATING AI INTO RADIO ELYSSA'S DIGITAL PRESERVATION STRATEGY

A significant portion of Tunisia's radio archives have been lost due to financial pressures resulting in the closure of many stations, and lacking resources many of the remaining collections are not being properly maintained. Radio Elyssa's sound archives are of great value in documenting Tunisian history and culture across the southern region. Unfortunately, Radio Elyssa faces new challenges that put their collections at risk: economic difficulties, gaps in cultural preservation capacity, and huge amounts of digital content to manage. The station broadcasts 24 hours a day, seven days a week, with one-hour programs recorded in an MP3 format at 128 kbps, totaling 1.2 GB of content per day of audio recording. In the past, these recordings were made on physical media (CD) and many of these recordings were lost during multiple location moves between 2012 and 2014. Additionally, at some points the station was relying on YouTube as a backup format which is also not a sustainable archival medium. In 2020 they received a new radio license through the The Independent High Authority for Audiovisual Communication (HAICA) [1], and the law obliged them to adopt a preservation strategy

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

[2]. This partnership explores how AI can help in the development of a sustainable and affordable preservation strategy.

To ensure the long-term preservation of cultural heritage, a reliable preservation strategy includes the creation of geo-redundant preservation of 2 or 3 master copies, the migration to new formats and media, awareness of obsolescence, controls of integrity, accurate metadata, transparent rights management, professional monitoring, and adaptation of standards. Radio Elyssa's archives face digital preservation challenges due to technical, financial, and human constraints, which influence the choice of storage formats. They aim for an optimal balance between storage space and quality and implement regular verification, migration, and cloud backup procedures to minimize data loss.

While Radio Elyssa's archive is primarily comprised of audio recordings, it also includes image and video content used for promotion and education. They use lossless formats like JPEG and FLAC [3] for images and audio. The MP4 format with X264 encoding is chosen for video, specifying a parameter (CRF) that adapts the bit rate according to the content. A value between 15 and 20 offers a compromise between quality and storage space.

Storage solutions include multiple SSDs for backup with a plan to migrate to higher quality storage media every five years. Data integrity is verified prior to migration. Initial investments have covered the purchase of storage media and verification procedures but the long-term plan is for an extensive cloud storage service to reduce the need for local storage.

Innovation in AI can improve archive management and has the potential to quickly analyze large data sets and automate archiving tasks. The Radio Elyssa case study focuses on the judicious use of targeted AI applications to optimize specific tasks while adhering to archiving best practices and increasing audience engagement. The goal of implementing AI is to reduce costs, including technical training, and to improve efficiency. The partnership between DCAT and Radio Elyssa tested the following AI tools to identify which ones would be most helpful for Radio Elyssa's long-term goals of preservation and audience engagement:

- Chatbots for video and sound migration
- AI for audience engagement through image enhancement
- AI for audio enhancement and music creation
- AI for transcription and subtitling of Tunisian Arabic
- AI for Analyzing and Correcting Video, Creating Subtitles, and Detecting Deepfakes

This article will explain how these tools were tested, and the benefits and challenges that arose. It is important to note that AI must be carefully implemented to complement human monitoring that determines appropriate digital preservation methods.

III. Testing Chatbots for Video and Audio File Migration

As part of long-term preservation and access, Radio Elyssa migrates its production-output audio and video files to higher quality formats. Audio MP3 files are converted to the lossless format FLAC, and video MP4 files with h264 codec are converted to a VP9 codec [4] (with a CRF of 18) which supports higher resolutions up to 8K. Chatbots [5] such as ChatGPT [6] can simplify this process of audiovisual content migration. For those unfamiliar with Python tools for audio-visual [7] tasks or FFmpeg [8], the use of chatbots can help guide the migration process. Chatbot technology can help break down the steps of all parts of the migration process, enabling real-time feedback, simplifying progress tracking, and making adjustments as needed.

The team from Radio Elyssa and DCAT tested four chatbots (Chat GPT, You.com, Caktus AI and writesonic.com) for use in automating migration tasks. They found that all the chatbots could help automate their tasks, although Caktus.AI offers more parameterization and customization of the conversion. Recognizing that automating large processes would be expensive with a paid solution, they noted that Chat GPT and you.com's free services would be preferred for budget-conscious archives. They also noted that given limited technical knowledge on the part of archivists and the potential to make mistakes when using chatbots, a better solution may be the simple use of open-source app shutter encoder [9].

The chatbot tests raised many questions about the technical expertise, both of coding and of file formats and codecs, that is needed in archival management. Although using chatbots to automate tasks may seem accessible to staff with more programming experience, it still may be inaccessible to some staff, without significant investment in training. More research is needed to understand how to develop the human skills needed for archival tasks in a digital preservation strategy incorporating AI. On one hand, human oversight is needed when implementing AI tools, and on the other hand, staff must be properly trained to provide that oversight.

Fig 2: ChatGPT/ Youchat

Fig 3: Caktus AI / Writesonic

IV. AI for Audience Engagement Through Image Enhancement

Radio Elyssa's photographic archives, featuring radio shows, guests, and special events, represent valuable resources to engage audiences, albeit with somewhat limited quantity. These images are an effective means of attracting audience attention before and after broadcast through social media platforms such as Intsagram. Unfortunately, many of these images have been shared and stored in low TIFF quality versions, reducing their effectiveness to inform audiences about radio programming. For long-term preservation these images converted to JPEG, and using AI tools can help with the improvement, enhancement and optimization of images, making them more legible and suitable for use on different platforms.

Due to limited expertise working with images in the team, they chose to test tools that could perform simple tasks such as object and background removal with remove.bg and automatic color correction and super-resolution [10] with Real-ESRGAN which performs Real-World Blind Super-Resolution training with pure synthetic data [11]. Super-resolution creates an increase of resolution in image processing and can be very helpful in making images usable, but the increase in storage space raises questions about the feasibility of long-term storage. The tests also raised significant questions about the authenticity of images and concerns of how tools such as super resolution could change, for example, facial features in a photograph. The tests raised the important point that the responsibility for preserving the authenticity of archives rests with humans.

V. AI for Audio Enhancement and Music Creation

Making the Radio station's sound archive available for long-term preservation and re-use has a few challenges. Questions of copyright mean that not all music is accessible for re-use, and external recordings suffer from poor sound quality due to non-professional equipment and a lack of sound management expertise. To address these concerns, the team tested Artificial Intelligence tools for sound improvement, automatic mixing/mastering, and music creation.

The team used Krisp.ai and Adobe podcasts [12] to test how they were able to enhance and automatically attenuate noises in external interviews, in order to save time in the studio. They found that Adobe was able to reduce noises by 100%, while Krisp.ai was not able to reduce all the noises, despite the wide options that it offers. For automated mixing/mastering algorithms they tested Mixcord and Square and found that they improved consistency and productivity, mixing 100 times faster than manually. They tested Jukebox [13], MuseNet [14], AIVA, and beatoven.ai to generate customized music in order to expand Radio Elyssa's musical offerings.

These tests provided important insights into the potential benefits and challenges for using AI in a radio sound production environment and archive. The team found that while these AI tools helped to produce high-quality programs, optimize editing time, and saving on archiving costs human supervision is still essential. They found, for example, that AI algorithms struggle with complex audio, requiring human adjustments. They noted that because sound design choices ultimately shape Radio Elyssa's programming character and impact, they need to be controlled by experienced

professionals. Additionally, the use of AI-generated music raised questions about the need for human verification to ensure copyright compliance. Additionally, using AI generated music can be problematic because listeners are not accustomed to it.

In summary, the targeted, responsible AI use for specific audio tasks can optimize Radio Elyssa's post-production processes. But human expertise, judgment, and curation remain essential to ensuring the audio heritage's artistic quality and uniqueness. The tests revealed the important point that AI augments - rather than replaces - human capabilities.

VI. AI for Transcription of the Tunisian Arabic Dialect

Transcription of speech is a powerful tool that many archives employ for increased access to their collections. Speech-to-text tools can produce transcripts for search and research purposes, and can help create subtitles for video content. Radio Elyssa and DCAT tested multiple speech-to-text tools including Kaptioned, Kapwing, Free subtile.ai and Subtitlebee and found that none of them could sufficiently recognize the Tunisian Arabic dialect. Despite the availability of parameters for the Tunisian dialect in different systems and programs, support remains limited, leading to unacceptable transcription errors.

One possible solution could be the development of speech recognition software or machine learning models specifically trained on Tunisian dialects in collaboration with archives. This would improve the accuracy of the transcription process and better support for the language and its nuances. Another approach could be to use local linguists and experts to manually transcribe and annotate audio recordings, to create a training dataset for machine learning models. This dataset could be used to improve existing automatic transcription systems or to develop new ones according to the specific needs of the dialect. A combination of technology and human expertise is likely to be required to effectively address the challenges posed by the Tunisian dialect. This will ensure the accuracy and reliability of the text produced from these tools.

VII. AI for Correcting Video, Creating Subtitles, and Detecting Deepfakes

Although the archives of Radio Elyssa contain only a limited number of videos (programs, interviews, news), AI offers the potential to help improve these images with tools such as motion detection, green screens, or background removal (runway ml). Text-to-video tools [15] such as Gen2 can delete objects or create video from text, opening up new potential for post-production or creative re-use of video from the archives.

AI tools offer a lot of potential also for creating subtitles and translations on video files, but the problems of automatic generation of subtitles remains a challenge for the Tunisian Arabic dialect, as explained in the section about transcription.

Finally, Radio Elyssa and DCAT have been testing tools to detect deepfakes [16], such as Deep AI's DeepFake-o-meter. The team understands that broadcasters and archives need to take careful steps to ensure the authenticity of content and to avoid accidentally sharing or storing deepfakes in the archive. Unfortunately, the development of deepfake detection tools helps to aid the development of deepfakes themselves, and the speed of evolution of deepfakes makes it impossible to create tools to detect them. The team is aware that deepfakes are a serious concern to the ethical responsibilities of broadcasters and archives, and that humans must play a role in carefully checking sources in order to verify authenticity.

VIII. Conclusion

Radio Elyssa and Digital Cooperation Association Tunisia's partnership created a rich environment for exploring the potential of AI to assist the radio station's long-term digital preservation strategy. Given that the radio station's goal is production, it is helpful to partner with an outside organization that is oriented towards the technical and intellectual demands and challenges of preservation. Radio Elyssa is facing technical and financial difficulties, including a lack of resources and expertise and creative solutions and the use of AI have helped the station to reduce the costs of technology, equipment and training in order to fulfill the mission of preserving Tunisia's cultural heritage, while respecting ethics and rights. However, challenges remain with regards to the accuracy of

audio transcription, compliance with copyright for AI-generated content, and human responsibility for the ethical preservation of cultural heritage over the long term.

Continuing professional training of employees mixed with the development of strong digital preservation governance is essential to ensuring Radio Elyssa's digital sustainability. A balanced combination of AI software, open-source software, such as cup cat, and shutter encoder, and human know-how prevents over-dependence on either. Responsible use of AI means that it must be supervised and controlled by human experts at all stages. Humans must define policies, understand the limitations of AI, and make the final decisions. Preserving high quality audiovisual heritage depends on informed human decisions supported by tools used with discernment. By investing responsibly in infrastructure and procedures that balance constraints and best practices, Radio Elyssa's archives can be sustainability preserved. Through this partnership with DCAT, Radio Elyssa recognized that careful adoption of AI tools, with appropriate governance structures and human supervision, will be essential to ensure ethical [17] and sustainable cultural preservation.

1. REFERENCES

[1] "Présentation", HAICA. [Online]. Disponible: https://haica.tn/presentation/

[2] "Digital Preservation Planning", IASA - Technical Committee for Audiovisual Archives. [Online]. Disponible: https://www.iasa-web.org/tc04/digital-preservation-planning

[3] L. Katz, "How to get the most out of your FLAC files: What's so special about the free lossless audio codec?," SoundGuys, Jun. 8, 2023. [Online]. Available: https://www.soundguys.com/flac-audio-guide-28859

[4] "VP9 Codec: Google's Open Source Technology Explained," wowza.com, Mar. 2021. [Online]. Available: https://www.wowza.com/blog/vp9-codec-googles-open-source-technology-explained.

[5] D. Noever and K. Williams, "Chatbots as Fluent Polyglots: Revisiting Breakthrough Code Snippets," arXiv preprint arXiv:2301.03373, pp. 1-20, 2023.

[6] K. Lehnert, "AI Insights into Theoretical Physics and the Swampland Program: A Journey Through the Cosmos with ChatGPT," arXiv preprint arXiv:2301.08155, pp. 1-10, 2023.

[7] "AV Python Carpentry," Amia Open Source, 2021, [Online]. Available: https://amiaopensource.github.io/av-python-carpentry/index.html.

[8] H. Zeng, Z. Zhang, and L. Shi, "Research and Implementation of Video Codec Based on FFmpeg," 2016 International Conference on Network and Information Systems for Computers (ICNISC), Wuhan, China, pp. 184-188, 2016, doi: 10.1109/ICNISC.2016.049.

[9] Shutter Encoder, Documentation, available: https://www.shutterencoder.com/documentation.html.

[10] Y. Romano, J. Isidoro, and P. Milanfar, "RAISR: Rapid and Accurate Image Super Resolution," IEEE Transactions on Computational Imaging, vol. 3, no. 1, pp. 110-125, 2016.

[11] X. Wang, et al., "Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[12] V. Ayeni, "Why You Shouldn't Use Adobe's Enhanced Speech AI Tool for Voiceover Production," APVA, 01/10/2023. [En ligne]. Disponible: https://apva.africa/why-you-shouldnt-use-adobes-enhanced-speech-ai-tool-for-voiceover-production/.

[13] P. Dhariwal, et al., "Jukebox: A generative model for music," arXiv preprint arXiv:2005.00341, 2020.

[14] Pal, Abhilash, et al. "Musenet : Music Generation using Abstractive and Generative Methods"International Journal of Innovative Technology and Exploring Engineering, vol 9,iss 6 ,Apr 2020.

[15] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," arXiv:2302.03011 [cs], Feb. 2023.

[16] K. Zhang, J. Yang, X. Tang and J. Luo, "Deep Learning for Deepfake Detection: A Comprehensive Review," arXiv:2202.06095 [cs.CV], Feb. 2022.

[17] UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. SHS/BIO/PI/2021/1.

Digital records curation at the East and Southern African Universities Institutional Repositories (IRs)

Juliet Erima
Moi University
Kenya
julieterima@gmail.com

Tshepho L. Mosweu
University of Botswana
Botswana
lydhoss@gmail.com

Abstract - Most of the knowledge generated in academic institutions today is in digital form. Given that institutional repositories (IRs) across universities receive, preserve and make access to digital assets. The aim of this study is to assess the status of digital curation at Institutional repositories in selected Universities in Botswana and Kenya. The study takes a quantitative approach whereby data was collected through survey questionnaires administered amongst university IR staff in Botswana and Kenya. The data collected was analysed and presented with tables and figures. The Open Archival Information System (OAIS) Functional Model was used in this study as a lens to investigate the problem. The findings of the study show that public universities in Botswana and Kenya have established IRs that ingest digital records into their custody. Most resources ingested include thesis and publications by academic staff and students. These IRs store their digital records on local servers and other storages like CDs. This study found that the majority of the IRs both in Botswana and Kenya do not have digital records preservation plans. This study recommends the use of the OAIS model to preserve, manage and make access to digital records at East and Southern African public Universities Institutional Repositories.

Keywords: Botswana, Kenya, Universities, Digital r ecords curation, Institutional repositories

INTRODUCTION

Universities and higher education institutions are in the business of generating a lot of information and knowledge resources, both in analogue and digital formats. By comparison, digital content has become increasingly ubiquitous in present day organisations. Anderson and Rainie (2012) acknowledge that "we swim in a sea of data [...] and the sea level is rising rapidly". Institutions are increasingly finding themselves "between a rock and a hard place" when facing rapidly changing technologies and the sheer volume of digital creation (Hedstrom, 1998). Due to the exponential creation of born-digital materials, information is being lost nearly as soon as digital assets are produced. As a result of this, individuals, institutions, and society as a whole need an accurate, complete and usable record of human activities, and an appropriate legal and institutional framework in which to use that record. Without trustworthy records, people and institutions cannot make informed decisions, verify existing information, evaluate evidence, hold others accountable, construct accurate histories or develop new knowledge (Prom, 2011). An authentic record does not preserve itself, and even the best-intentioned record creators often lack the resources or expertise to act as permanent custodians for non-current records. Nor can we rely on those who provide the service of temporarily storing and transmitting records to permanently preserve an interpretable record of human activity (Prom, 2011).

Harvey (2010) posits that technical obsolescence or fragility, lack of resources, ignorance of good practices, and uncertainty over appropriate infrastructure – all constitute serious risks to data. In previous years, digital preservation efforts originally focused on ensuring that material survived technical obsolescence and organisational mismanagement. Preservation implied a passive state, where material would be "dumped" in an inaccessible "dark archive", with only a few authorised users, to ensure that it retained its integrity and authenticity.

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

Lately, the focus has shifted to ensuring that digital material is managed throughout its lifecycle so that it remains accessible to those who need to use it. Metadata is used to both improve accessibility and discoverability; and to control authentication procedures, creating audit trails to ensure that material cannot be accessed or altered by those not authorised to do so. Digital material is actively preserved, used and reused for new purposes, creating new materials. Unfortunately, relatively few institutional repositories in African public universities have implemented systematic institutional functions to preserve digital records in their keeping. Institutional repositories need a practical method to capture, preserve and provide access to records like email, blogs, digital photographs and unpublished reports, which are at extreme risk of loss over the medium and long term (Prom, 2011).

According to Walters and Skinner (2011), the responsibility for the custody and preservation of cultural heritage lies squarely upon the shoulders of librarians and archivists. This paper assesses the status of digital curation at Institutional repositories in selected Universities in Botswana and Kenya. An Institutional repository (IR) has been defined as a library of digital objects and associated metadata from a single institution (Clobridge, 2010)

Research Problem

Universities and other research organizations create and amass large volumes of digital assets and information which include administrative records, theses and dissertations, research publications, multimedia collections, digital surrogates of cultural material, learning objects, course materials, among others (Schmidt, Ghering and Nicholson 2011). Tindermans (2009) addressed the subject of digital preservation in the community pointing out that the huge volume of digital content, diverse variety of digital objects formats coupled with rapid technological changes that gave rise to an influx of new versions was a red flag that could not be ignored. Institutional repositories in many public universities in Africa such as Botswana and Kenya lack comprehensive, campus-wide digital preservation programmes or guidelines. Intentional digital preservation strategies are necessary in order to respond to the increase in digital content - especially in technology-dependent formats - and to provide prolonged access to digital records and archives. The goal of this research study is therefore to determine what is occurring in institutional repositories of selected universities in Botswana and Kenya with regard to digital records curation and to eventually propose a strategy that can be adopted by these institutions to support the long-term preservation and access of digital records and archives.

Research Objectives

The objectives of this paper are as follows:

1. To establish how IRs in selected universities acquire digital records
2. To evaluate the methods used to store digital records in the IRs
3. To investigate how digital records are managed in the IRs
4. To establish the preservation strategies for digital records in the IRs
5. To find out the procedures for access and use of digital records in the IRs
6. Propose recommendations to enhance digital curation practices in the IRs.

The Concept of Digital Curation

The term digital curation was first used in 2001 to refer to digital preservation, data curation, and the management of assets over their lifecycle (Yakel, 2007). Today, the term digital curation is increasingly being used for the actions needed to add value to and maintain these digital assets over time, for current and future generations of users (Beagrie, 2008). According to Yakel (2007), “Digital Curation is the active involvement of information professionals in the management, including the preservation, of digital data for future use”. The Digital Curation Centre (2020) defines digital curation as “maintaining and adding value to digital research data for current and future use” and adds that “it encompasses the active management of data throughout the research lifecycle”. According to Yakel (2007), “Digital Curation is the active involvement of information professionals in the management, including the preservation, of digital data for future use.” Given the diversity of its stakeholders and of the environments in which it is conducted, digital curation potentially involves anyone who interacts with digital information during its lifecycle.

For purposes of this study, digital curation shall be defined as the active involvement in the management, including the preservation, of digital resources for future use. This intentionally broad definition is slightly adapted from Yakel. It omits the restriction to who is involved and uses the term "digital resources". Note that the focus on future use can be a very close or a very distant future. Ball (2010) defines digital curation by stating that digital curation in IRs must be seen and understood together with terms of preservation and archiving.

In almost all areas of society, but in particular in science, research, and scholarship, the ability to effectively create, share and use digital resources has risen to form a crucial ability. The ability to manage these assets for current and future use is equally critical for a sustainable society. Institutional repositories play a crucial role in the preservation and making access to digital data and records through IRs. A study by Kakai, Musoke, and Okello-Obura (2018) found that libraries at Universities in East Africa were taking the lead in initiating and implementing IRs.

Models And Standards of Digital Archives Curation

In 2008, Higgins proposed a lifecycle in seven phases, namely the Digital Curation Centre (DCC) curation lifecycle model (2008), based on Pennock's (2007) lifecycle approach to digital curation. This lifecycle is composed of the following phases: create or receive; appraise and select; ingest; preservation action; store; access, use and re-use; and finally, transform, which links back to the first phase. According to Higgins:

"This lifecycle approach ensures that all the required stages are identified and planned, and necessary actions implemented, in the correct sequence. This can ensure the maintenance of authenticity, reliability, integrity and usability of digital material." (Higgins, 2008).

National Archives of Australia (2006) opines that Intellectual and physical management systems that are employed to store, manage, retrieve and deliver digital objects should, ideally, be based on open standards to ensure sustainability of the systems over time. Open standards exist for format types, for operating systems, disk drives and so on. If proprietary systems are used, digital objects could be lost or rendered uninterpretable over time. The Archives Domain is advocating that digital archiving solutions be based on open standards such as the Open Archival Information System (OAIS) Reference Model ('Blue Book' digital preservation framework – ISO 14721: 2003).

Theoretical Framework

There are different models that may be used in the management of records such as the Records Life-Cycle Model, the Records Continuum Model and the OAIS Model. However, this study uses the Open Archival Information System (OAIS) Functional Model as a lens to investigate the status of the preservation and access of digital records by the public universities in Botswana and Kenya. The OAIS model categorizes the core set of tools with which an OAIS-type archive meets its primary mission of long-term preservation of information and access by the users (Digital Preservation Coalition 2015). Figure 1 depicts the OAIS model.

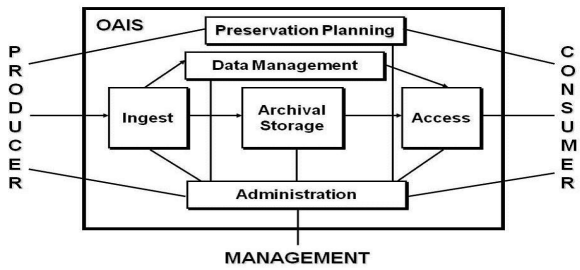


Figure 1 The OAIS Functional model. Source: Digital Preservation Coalition 2014)

The adoption of OAIS was purposely for its wide applicability for long-term preservation to any context, but principally in a digital environment, hence its relevance to the present study. The model is also suited for application in organizational and institutional set-ups such as public universities.

Methodology

This study used the quantitative approach whereby data was collected through questionnaires administered amongst university institutional repositories staff in Botswana and Kenya. The researchers desired only one response from each institution, preferably the staff in charge of the IRs. Online survey questionnaire was sent to four (4) public universities in Botswana and Kenya, giving a total of eight (8) questionnaires. The survey did not include private universities, colleges, or vocational training institutions. In total 8 responses were received. The data collected was analysed and presented in tables and figures.

Results And Discussions

The following section presents the results as per themes drawn from the research questions of this paper which are: acquisition of digital records in the IRs, management of digital records in the IRs, preservation strategies for digital records in the IRs, access and use of digital records in the IRs as well as recommendations to enhance digital curation practices at the IRs.

Ingest Of Digital Records

The ingest function as per the OAIS functional model relates to the receipt of information from sources, its packaging, acceptance of a Submission of Information Package (SIP), verification and the transfer of the created Archival Information Package (AIP) to the archival storage.

Types of digital records - Some of the digital records received by IRs include e-prints (both pre- and post-prints), grey literature (especially e-theses), working papers, technical reports, books and book chapters, conference papers, posters and administrative records (Jones, n.d). Respondents were asked to identify the types of digital records they received at their Institutional repositories. As shown in Figure 1, the majority of the respondents (42.9%) indicated that they receive thesis at their repositories followed by scientific research papers, artefacts, research publications, peer reviewed and published prints all at 14.3% each. Public universities generate large volumes of digital content emanating from three broad activities namely teaching, research and extension and outreach.

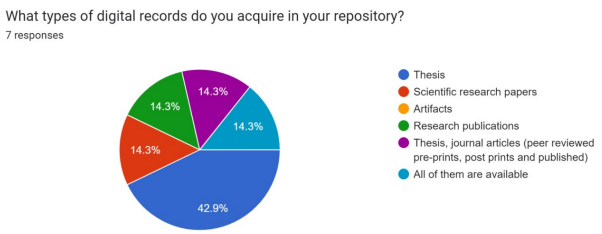


Figure 2 Types of Digital Records

Source of digital information - The study also sought to find out the source of digital records by IRs. Figure 3 shows that academic staff was the most cited source of digital records at 37%, other sources included students at 25%, administrative staff at 25% while the rest (publishers, postgraduate students and academic staff, students and administrative staff) stood at 12.5 % each. A study by Kakai (2018) revealed that lack of open access policies operating within institutions and lack of awareness of open access IRs among researchers and academicians were some of the factors that contribute towards limited acquisitions.

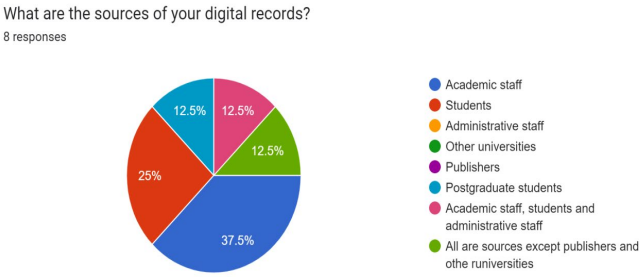


Figure 3 Sources of Digital Records

Best practice in digital archiving demands that archival repositories should formulate and implement collection development policies addressing the materials that the archives retain and what that which is not collected (Noonan and Chute 2014). The study findings indicated that majority of the institutions (75%) had collection development policies while two of the institutions were in the process of developing such documents.

Information Attached to the Digital Content

When asked about the essential information that should be attached to the digital content before acceptance into the IR, respondents answered as follows:

R1: Plagiarism similarity check report and author consent form.
R2: Delivery list from postgraduate or IR submission Form.
R3: It's provenance.
R4: For hardcopy publication you have to scan it to digitize and soft.
R5: Item description (author, title, publisher, citation etc.).
R6: Thirteen elements from Dublin Core metadata Standard.
R7: ORCHID ID is critical

Digital records verification - OAIS functional model also requires that information be verified during the ingest function. Exlibris Knowledge Centre (2022) is of the view that the responsibility for the quality and accuracy of Institutional Repository

content belongs to the source of data. The study respondents were asked to indicate how the information they received was verified. The R1 indicated that they have an office designated to repository administrator who is responsible for verification; R2 indicated that they have a Correction of Thesis form; R3 revealed that they do not verify information yet; R4 revealed that they use a Sherpa Romeo; R5 said that they compare the information with the physical document. One respondent did not answer this question; R6 indicated that metadata is verified by the Repository manager before the content can be published while R7 said metadata is verified through the registry of researchers.

Packaging Of Digital Information

The study sought to find out how the digital information received was packaged. Two of the respondents did not answer this question; however, the rest of the respondents gave the following answers:

R1 When the digital content is received at the office of repository administrator, its first run through "Turnitin" the anti-plagiarism software to verify the level of plagiarism whether it is within the University's accepted standard. Secondly, it's processed by classifying to determine the repository community and subject which is treated in the document. Finally, the record is entered to the IR.

R2 Once uploaded, the work can be searched via author, subject, title etc.

R3 We have not yet received digital information, only print.

R4 For hardcopy publications you have to digitize by scanning.

R5 Information is arranged into groups called communities which are subject-specific. In the case of [University X] the communities have similar names as university Faculties. So basically, the information is packaged according to faculties.

R6 The repository is made up of different communities within the University.

Storage Of Digital Records by IRs

Archival Storage function is about the storage, maintenance, and retrieval of archival information packages (AIPs). When asked how they store digital records, the majority of the respondents (85.7%) indicated that they store their records on servers while only 14.3% indicated CDs and hard copies as depicted in Figure 4. None of the respondents indicated that they store their records on either clouds or servers. One respondent did not attempt this question.

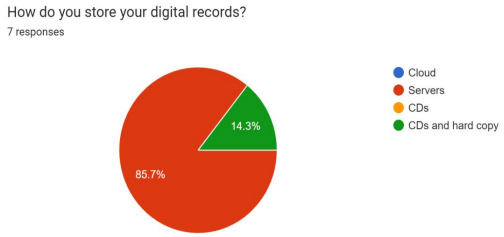


Figure 4: Storage of digital records

Management Of Data

Archival Storage function is about the storage, maintenance, and retrieval of archival information packages (AIPs). It accepts AIPs submitted from the ingest function, assigns them to long term storage, migrates AIPs as needed, checks for errors, and provides requested AIPs to the Access function. Some University IRs in this study stated that they use the DSpace software; however, Kakai et al (2018) argue that software is not easy to install and maintain.

Query requests - The respondents were asked to state the procedure for executing query requests and generating results. The responses were as follows:

R1 The users have been assigned to a specific email that receive and answer users' questions.

R2 Via manual or online request through email and the same for results.

R3 Not yet applicable [IR not yet established]

R4 Searching using Author, title and subject.

R5 Contact the IR Manager.

R6 The D-Space platform sends emails to administrators and if there is a query then the admins will address it.

R7 DSpace's JSPUI. The JSPUI defines several filters, listeners and servlets to process a request.

Reports generation - The respondents indicated that they generated the following reports: deposits reports and entries, usage statistics (downloads, views), most popular items and authors, content statistics, search statistics, storage statistics, statistics by country, items added in a given certain of time, total items in a repository

Preservation Planning

Preservation planning function supports all activities meant for long term preservation and accessibility of digital records.

Preservation tasks - Respondents were asked to state whether they had preservation or migration plans in place and the majority of them (71.4%) indicated that they do have preservation plans, while 28.6% indicated that they do not have preservation plans as depicted in Figure 5.

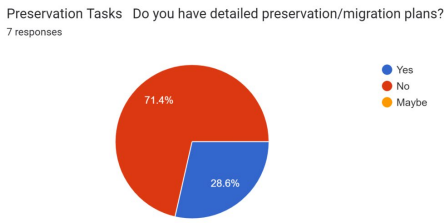


Figure 5 Preservation Tasks

Evaluation and risk analysis of content - When asked how often they do evaluation and risk analysis of content, 80% of the respondents indicated that they do not do any risk analysis of content or they don't know about it, while only 20% indicated that they do it daily and quarterly.

Access And Use of Digital Records

The Access function relates to the user interface that allows users to retrieve information from the archive on request. Kakai, Musoke and Okello-Obura (2018) argue that in the digital environment, library users are interested in easily accessing full-text information resources, and these should be readily available from IRs. Respondents in this study were asked to comment on how user- friendly their interface was based on a scale of 1-5 where 1 was fairly friendly and 5 very friendly. The majority of them (50%) chose 5, 33.3 chose 2 while 16.7 chose 1 as depicted by Figure 6.

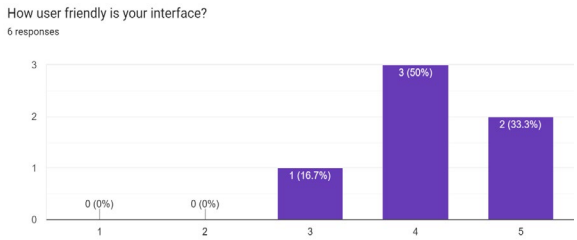


Figure 6 User friendliness of the interface

To explain further responses in on the user friendliness of the IRs interfaces as presented in Figure 5, the respondents had this to say:

R1 The terminologies used are common English and can easily be understood by anybody who understands English.
R2 It easy to retrieve a document by author or title.
R3 Not applicable [Yet to establish an IR].
R4 It is easy to navigate.
R5 It provides greater visibility and accessibility at all times.
R6 It is easy to navigate since there are no pictures that can distract the user, less customization.
R7 All features are clear.

Recommendations To Enhance Digital Curation Practices

Based on the findings from the data collected, this paper presents the following recommendations:

Expansion of digital information content that is received by IRs to other materials such as digital archival materials and photographs

Universities should diversify their sources of digital records to ensure that more information is preserved and to comply with the legal deposit legislation. Other sources which are not target for the IRs surveyed in this study include other universities and publishing houses.

Cloud storage is recommended with its capacity to improve access to sharing of information and its preservation.

Implementation of the OAIS model for digital records preservation by IRs as it promotes long term preservation of digital records and may allow of interoperability with other IRs.

Conclusion

This study has established that some universities in Botswana and Kenya do have IRs that ingest digital records into their custody though the resources are mostly limited to thesis and publications by academic staff and students. The study also found that the storage of digital records by IRs in Botswana and Kenya is mostly on local servers and CDs. Furthermore, the majority of the IRs surveyed indicated that they do not have preservation plans for their digital records. As digital records are increasingly being generated, IRs would play a crucial role in the preservation of digital records in the Southern and Eastern countries such as Botswana and Kenya. The current survey may serve as the basis for bigger research to include more IRs in Southern and Eastern Africa.

1. REFERENCES

[1] Anderson, J. and Rainie, L. (2012). The future of the Internet. Washington, DC: Pew Research Center.

[2] Ball, A. 2010. Review of the state of the art of the digital curation of research data. Bath, UK: University of Bath.

[3] Beagrie, N. (2008). Digital curation for science, digital libraries, and individuals. International Journal of Digital Curation, 1(1), 3-16.

[4] Clobridge, A. 2010. Introduction in Ruth Rikowski. Building a Digital Repository Program with Limited Resources. Chandos Publishing (Oxford) Limited. UK.

[5] Digital Curation Centre (DCC). 2020. What is digital curation? Available at: http://www.dcc.ac.uk/about-us [Accessed 28 February 2020].

[6] Digital Curation Centre. 2008. DCC curation lifecycle model. Available at: http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf

[7] [Accessed 7 February 2019].

[8] Digital Preservation Coalition. 2015. Digital preservation handbook. 2nd ed. Glasgow: University of Glasgow.

[9] Exlibris Knowledge Centre. Institutional Repositories in CDI. Available online at https://knowledge.exlibrisgroup.com/Primo/Content_Corner/Central_Discovery_Index/Knowledge_Articles/Institutional_Repositories_in_CDI (Accessed 22 January 2023).

[10] Harvey, R. (2010). Digital Curation: A How to Do It Manual. New York: Neal Schuman.

[11] Hedstrom, M. (1998), "Digital preservation: a time bomb for digital libraries". Computers and the Humanities Vol. 31: 189-202.

[12] Higgins, S. 2008. The Digital Curation Centre Lifecycle Model. The International Journal of Digital Curation 3(1): 134-140. Available at: http://www.ijdc.net/article/download/69/48/0 [Accessed 24 March 2019].

[13] Kakai,M., Musoke, M.G.N and Okello-Obura, C. 2018. Open access institutional repositories in universities in East Africa. Information and Learning Science. Vol. 119 No. 11, 2018 pp. 667-681.

[14] National Archives of Australia. 2006. Designing Information and Recordkeeping Systems (DIRKS): a strategic approach to managing business information. Canberra: National Archives of Australia.

[15] Noonan, T. and Chute, D. 2014. Data Curation and the University Archives. The American Archivist, Vol. 77, No. 1 (SPRING/SUMMER 2014), pp. 201-240.

[16] Pennock, M. 2007. Digital curation: a life-cycle approach to managing and preserving usable digital information. Journal of Library and Archives 1: 12-24.

[17] Prom, C. 2011. Making Digital Curation a Systematic Institutional Function the International Journal of Digital Curation. Vol 6(1): 139-152.

[18] Schmidt, L., Ghering, C. and Nicholas, S. 2011. Digital curation planning at Michigan State University. Association of Library and Technical Services 55(22): 104-118.

[19] The Open Archival Information System (OAIS) Reference Model: Introductory Guide. 2014. The Digital Preservation Coalition. Great Britain (2nd Edition).

[20] Walters, T. and Skinner, K. 2011. Digital curation for preservation. Washington, DC: Association of Research Libraries.

[21] Yakel, E. (2007). Digital curation. OCLS Systems and Services Vol. 23 No. 4, pp. 338-339.

From Silos to Community

The Path to a Holistic Digital Preservation Policy

Laura McCann
New York University
USA
lm103@nyu.edu
0000-0001-7821-3127

Weatherly A. Stephan
New York University
USA
was227@nyu.edu
0000-0002-6381-2036

Abstract – While New York University Libraries has a long history of and commitment to digital collecting and preservation efforts, the institution did not have any policies governing the services and activities of digital preservation prior to 2022. This paper details the creation of a holistic digital preservation policy statement, with contributors from across ten functional units at NYU Libraries. The policy was grounded in the Libraries' mission and values–including deep commitments to inclusion, diversity, belonging, equity, and accessibility–and drew on themes crafted by all members of the group to ensure their work was represented in the statement. The success of the policy group was rooted in its intentional formation and processes that acknowledged the distributed nature of digital preservation and emphasized the creation of a community of practice. Further, it laid the foundation for a more complete suite of preservation policies and forward-looking conversations about how to enact ethical and sustainable stewardship in digital collecting, access, and preservation practices.

Keywords – Community of Practice, Documentation, Policy, Preservation Strategy, Stewardship

Conference Topics – We're All in this Together; From Theory to Practice

I. Introduction

New York University (NYU) Libraries has had a long, deep involvement in the development of digital collecting and preservation practices. But despite this history and strong institutional commitment, NYU Libraries did not have any policies governing digital preservation prior to 2022. This paper describes the process of developing NYU Libraries' first digital preservation policy, which required the breaking down of silos to create a community of practice. Throughout the creation of the policy, the authors recognized the necessity of the distribution of digital preservation throughout the institution, along with the importance of centering ethical and sustainable stewardship practices in our digital-focused work.

From contributing to the development of the Metadata Encoding & Transmission Standard; adopting web archiving for special collections repositories in 2007 to the IMLS grant *Saving Data Journalism* to archive dynamic websites in 2019; launching international postcustodial projects such as the Afghan Digital Library and Arabic Collections Online; the publication of profession-wide standards such as "Digitizing Video for Long-Term Preservation" and "Guidelines for Preserving New Forms of Scholarship"; much of NYU Libraries' work in digital preservation has been in public, collaborative projects, often supported by grant funding [1-3]. Beyond the in-house research and development department, Digital Library Technology Services, these projects germinated across the Libraries in Research & Research Services, the Barbara Goldsmith Preservation & Conservation Department (Preservation Department), and special collections repositories.

In addition to many public-facing projects, NYU Libraries developed its own digital repository in 2011 and is actively engaged in the preservation of digital content, born-digital media, digitized analog content, and software. While many digital preservation practitioners at the Libraries recognized the importance of collaboration, there was also a tendency to work solely on localized decisions and technical frameworks without looking toward holistic needs across the institution.

II. Institutional Context

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

In the last five years, NYU Libraries has undergone a significant organizational change prompted by new leadership. In 2018, H. Austin Booth was appointed Dean of NYU's Division of Libraries, an organization that includes NYU Press and NYU-TV. Dean Booth reorganized the leadership, creating a flatter organizational structure that enables direct communication between the leaders of all functional groups. The Preservation Department then joined the senior leadership team, providing new opportunities for collaboration on digital preservation.

Following the reorganization, Dean Booth charged the leadership team with the creation of a digital library governance structure that focused on inclusive and equitable practices across digital collecting, preservation, and access in the Libraries. This aligned with one of the top strategic priorities for the organization–digital preservation–along with a growing recognition that areas of digital collecting were not open to all curators and collectors across the Libraries. For instance, when the digital library group was focused on supporting grant funded projects in the special collections repositories, there was no labor or resources to dedicate to new projects or collecting areas outside of grant structures. Furthermore, a reliance on grant funding for digital preservation projects created new work that was challenging to maintain after grant periods ended. This challenge prompted a commitment from the new leadership to the principle of ethical, sustainable stewardship of all collections, but especially when embarking on new digital collections work.

In 2020, senior leadership created a Digital Library Steering Committee to prioritize and resource proposed digital library projects from across the Libraries, inclusive of both general and special collections. The Steering Committee is informed by the work of a short-term Digital Library Selection Priorities Working Group, which produced criteria for digital collecting; and ongoing resource, labor, and workflow analyses provided by a Digital Preservation and Access Committee. The governance groups were all intentionally staffed with a balance of practitioners and curators to help inform each others' work and create a shared understanding of how digital collecting and preservation would evolve at the Libraries.

Into this new landscape, the lead author was promoted to Director of the Preservation Department in January 2021. While, as the former supervisor of preventive, general, and special collections conservation, McCann had established strong relationships with collections managers and subject librarians throughout the Libraries, digital preservation was new to her portfolio. Consulting with the second author, McCann realized that early-stage collaboration and building an inclusive community of practice would be critical to create policy for digital preservation at NYU Libraries [4].

1. III. Environmental Scan

In her first year as Director, McCann undertook both an external and internal environmental scan. Central in her learning about this area were concepts of the distributed nature of digital preservation from the literature, the work of the digital library governance groups, and, perhaps most importantly, informal connections with colleagues over a long tenure at NYU Libraries. An external review of policies showed diversity of style and scope at peer institutions. McCann's outreach to peer institutions also revealed that these policies were commonly authored by a single individual or single department, and many were focused solely on infrastructure. Considering NYU was already deeply engaged in the work of digital preservation, she determined it was most strategic to focus on a policy statement from which other policies could be developed.

Internally, in order to better understand the digital preservation touchpoints within NYU Libraries, McCann convened meetings with individuals and in small groups with colleagues in disparate departments. The internal environmental scan revealed that many colleagues are engaged in the work of digital preservation, and while there are some strong intra-organizational collaborations overall, other work was siloed. Many colleagues were surprised to learn that other departments were engaged in digital preservation. While this surprise in a few instances was attributed to an individual's narrow definition of digital preservation, usually it was due to the consequence of a large complex organization and the past hierarchical structure that hindered interdepartmental communications. For example, the Collection Management department regularly determines digital preservation terms on leased e-resources, and the Data Services

department had been ad-hoc managing purchased data files on hard drives: both opportunities for collaboration that did not lead to actual connections.

The environmental scan demonstrated the need for an inclusive and representational policy statement that would inform all digital preservation work at NYU Libraries. The lead author set about creating a policy task force by soliciting participants from all units represented on the senior leadership team, with the exception of the administrative units. The Dean and three Associate Deans were purposely excluded from the task force to provide more opportunity for colleagues at different levels in the organization.

The leaders of each unit were asked to nominate potential collaborators who were interested in the process and the goals of the group. While most nominations were for individuals within the reporting structure of the leader's department, there were notable exceptions. For example, the second author comes from a technical services department for archival collections, though on the policy task force she represented the curators in special collections repositories who acquire and appraise digital archives. The Libraries' Inclusion, Diversity, Belonging, Equity, and Accessibility (IDBEA) Steering Committee, whose co-chairs sit on the senior leadership team, was represented on the task force as well. In addition to the IDBEA Steering Committee and the Preservation Department, the task force had representatives from Collections & Content Strategy; Digital Library Technology Services; Libraries Information Technology Services; Knowledge Access & Resource Management Services; NYU Special Collections; NYU-TV; Research and Research Services; Scholarly Communications and Information Policy; Teaching, Learning, and Engagement; and User Experience.

2. IV. Policy Creation

The Digital Preservation Policy Task Force was convened in May 2022 and was charged with creating a concise digital preservation policy statement for the Division of Libraries within a five month period. The co-authors co-chaired the task force. The task force agreed to norms for anti-oppressive facilitation that are widely used at NYU Libraries, and adopted a participatory decision making model for work on the policy statement [5]. Over ten synchronous meetings with one to two weeks of asynchronous work between, the task force proceeded in three phases: research, drafting, and revision.

In the research phase, members shared resources that could inform the group's final product. Excerpts from Trevor Owens's *Theory and Craft of Digital* Preservation and the article "What's Wrong with Digital Stewardship?" formed the core of the group's initial reading [6-7]. The task force reviewed peer institution policies, ranging from public and private local and national institutions, such as Columbia University and the University of California, as well as model policies like the NASIG Model for Digital Preservation Policy. [8-10] Members also explored concepts adjacent to digital preservation work, such as maintenance and broken world vocabularies [11-12]. Throughout this process the group saw that the scope of digital preservation at academic research libraries goes beyond preserving and making accessible digital content to the broader work of helping researchers render their data preservable and reusable. Therefore, members determined that building digital preservation awareness and literacy must be part of the policy statement.

While reviewing the shared resources, each member took notes or highlighted salient points in a communal document, which then guided meeting discussion about the scope and shape of the policy statement. Drafting began with each member of the task force defining the term "digital preservation," surfacing assumptions and gaps in each others' knowledge. From this exercise, the task force defined specific themes that were then expanded into bullet points, with many of the themes taken directly from members' conversation in the shared readings discussion. The members also continued to outline the scope of what they saw as relevant for a policy statement while creating a narrative written policy from the bullet points.

The task force built a review of the completed draft by the senior leaders and department managers into the revision process, including gathering feedback from the colleagues at all levels in the organization who were engaged in digital preservation work. Once feedback was incorporated into the draft, the final version was copy edited before approval by the Dean. The completed policy

was circulated to the Libraries in the Dean's weekly newsletter, and published on the Libraries website in November 2022 [13]. The task force conducted an after-action review, highlighting that clear expectations, strong communication, and a compressed timeline provided helpful support for the creation of the statement. A deliberate representation of all areas of the Libraries also fostered new connections between colleagues and built greater understandings about the complexity of digital preservation needs for those who engage in this work daily.

3. V. Policy

The broad themes that the task force built out into the policy included a grounding in the Libraries' mission and values, open access, active and iterative maintenance, community outreach and collaboration, stewardship, external partners and tools, and challenges and risks. In writing the policy, the task force thought deliberately about how issues of inclusion and diversity could be represented in the statement, especially when considering the drive to collect digital materials from historically underrepresented communities. In the policy statement, the task force acknowledges the challenge in balancing the institution's commitment to accessibility to researchers with disabilities with the ethics of preserving materials that are restricted from use: "We commit to making material available to the widest possible range of users, including those with disabilities, and to adapting the process for making materials open and accessible as the work evolves. We make this commitment while recognizing that not all content may be made accessible" [13]. Another theme in the policy is the centering of people who do the work of digital preservation, particularly the work of maintenance that is frequently overlooked. Here the task force acknowledges that the work of digital preservation is dependent on having the resources to continually dedicate to the work: "We will meet the challenges of digital preservation head-on with the resources we have" [13].

The policy statement is defined internally as adaptable to changing priorities and needs, and as such, revision is expected: "We continuously evaluate our institutional approach, whether risk-tolerant or risk-averse, and adjust as necessary given the surrounding circumstances [13]." One area that will be considered in future revisions is to specifically call out the relationship of digital preservation and climate change, an area of increased interest and scholarship, as well as a priority for our University [14-16].

There is wide recognition from the task force and invested collaborators across the Libraries that the policy statement was a necessary foundation for the creation of a fuller suite of digital preservation policies. The policy statement provides guidance for other policies that we know to be gaps. These policies, including digital collection development and repository documentation, are crucial for meeting our commitment to ethical, sustainable stewardship of digital collections.

4. VI. Conclusion

With the policy live on the Libraries' website, the task force created a model for how to collaborate efficiently and grow a community of practice. The policy publication also instilled the practice of making institutional policies open and available on the Libraries' public-facing website. The Digital Preservation Policy Statement was the first of its kind to be published on the Libraries' policy page and spurred the publication of other foundational policies, such as the Open Metadata Policy [17]. Both of these policies provide users with email aliases to contact policy groups in an effort to broaden our community of practice.

While the policy statement was successfully launched and supported by the senior leadership and appreciated by many staff who work in digital preservation, the statement did not meet the expectations of some managers and practitioners of digital preservation. These colleagues voiced feedback that this policy statement did not address how digital preservation work is done day to day within departments. Instead, it is intended to provide both a foundation and document the institution's commitment to this work. Other feedback requested that archives be called out specifically. This feedback shows that we still have work to do to broaden our institutional understanding of what digital preservation work is across the Libraries, and reduce the bias toward archives and special collections when thinking about digital work.

We believe that the model for policy creation and growing a community of practice is replicable across

the profession, regardless of whether a change in leadership or organizational structure prompts the need. Deliberately engaging all colleagues across the Libraries; setting out with clear, achievable goals, then mapped into phases for the task force; and breaking down the actual authoring of the policy from capturing notes, definitions, and emergent ideas to bullet points to fully formed prose all contributed to our success. In addition, shared norms and alignment with institutional mission, values, and strategic goals helped both guide conversations and resolve areas of concern. Taking an iterative approach to policy creation ensures it is responsive to rapidly changing needs.

Finally, the policy statement, as well as the model of creating a digital preservation community of practice, provides an entry into challenging conversations about sustainability and ethical stewardship of born digital collections. Resources allocated to large digital collections with complex content must also be carefully considered, from curatorial decisions that are both time consuming and demand a comfort with risk, to the labor needed to accession massive born digital collections. Both as an organization and as professionals, we need to talk through these decisions, document our processes, and consider new opportunities.

5. REFERENCES

[1] P. De Stefano, et al., "Digitizing Video for Long-Term Preservation: An RFP Guide and Template," New York University Libraries, New York, NY, 2013 [Online]. Available: https://guides.nyu.edu/ld.php?content_id=24817650

[2] K. Boss, V. Steeves, R. Rampin, F. Chirigati, and B. Hoffman, "Saving data journalism: Using ReproZip-Web to capture dynamic websites for future reuse," in iPres 2019, LIS Scholarship Archive, September 2019. [Online]. Available: https://osf.io/preprints/lissa/khtdr/

[3] J. Greenberg, K. Hanson, and D. Verhoff, "Guidelines for Preserving New Forms of Scholarship," New York University Libraries, New York, NY, 2021 [Online]. Available: https://doi.org/10.33682/221c-b2xj

[4] D. Handel and M.A. Matienzo, "Facilitating and Illuminating Emergent Futures for Archival Discovery and Delivery: The Final Report of the Lighting the Way Project," Stanford University Libraries, Stanford, CA, 2021 [Online]. Available: https://doi.org/10.25740/jm302fq5311

[5] Anti-Oppression Resource and Training Alliance, "Anti-Oppressive Facilitation for Democratic Process," June 2017.

[6] T. Owens, Theory and Craft of Digital Preservation. Baltimore: Johns Hopkins University Press, 2018.

[7] K. Blumenthal, P. Griesinger, J. Kim, S. Peltzman, and V. Steeves, "What's wrong with digital stewardship: Evaluating the organization of digital preservation programs from practitioners' perspectives," Journal of Contemporary Archival Studies, vol. 7, article 13, 2020 [Online]. Available: https://elischolar.library.yale.edu/jcas/vol7/iss1/13

[8] Preservation Division, "Policy for Preservation of Digital Resources," Columbia University Libraries, New York, NY [Online]. Available: https://library.columbia.edu/services/preservation/dlpolicy.html

[9] E. A. Smith, J. Chodacki, M. Elings, T. Grappone, G. Janée, C. Macquarie, et al, "UC Digital Preservation Strategy Working Group: Phase One Report," UC Office of the President, University of California Systemwide Libraries, 2020 [Online]. Available: https://escholarship.org/uc/item/80v318pm

[10] NASIG Preservation Committee, "NASIG Model Digital Preservation Policy," NASIG, West Seneca, NY, 2002 [Online]. Available: https://nasig.org/NASIG-model-digital-preservation-policy

[11] Information Maintainers, "Information Maintenance as a Practice of Care: An Invitation to Reflect and Share," The Maintainers, 2021 [Online]. Available: https://themaintainers.org/information-maintenance-as-a-practice-of-care-an-invitation-to-reflect-and-share/

[12] D. Lovins and D. Hillman, "Broken-world vocabularies," D-Lib Magazine, vol. 23, no. 3-4, March-April 2017 [Online]. Available: https://www.dlib.org/dlib/march17/lovins/03lovins.html

[13] Digital Preservation Policy Task Force, "Digital Preservation Policy," New York University Libraries, New York, NY, 2022 [Online]. Available: https://library.nyu.edu/about/policies/digital-preservation-policy/

[14] B. Goldman, "It's not easy being Green(e): Digital preservation in the age of climate change," in Archival Values: Essays in Honor of Mark Greene, PennState ScholarSphere, November 2020 [Online]. Available: https://scholarsphere.psu.edu/resources/381e68bf-c199-4786-ae61-671aede4e041

[15] Office of Sustainability, "Sustainability," New York University, New York, NY 202x [Online]. Available: https://www.nyu.edu/life/sustainability.html

[16] E. Tansey, "Climate change, archives, and digital preservation," DPOE-N Workshop, April 21, 2023 [Online]. Available: https://docs.google.com/presentation/d/1E9lXy3Si2jHzKAM7v1NhHCYuAVHQOR6_xu5tYbB1Btl/edit#slide=id.p

[17] Metadata Policy & Implementation Committee, "Open Metadata Policy," New York University Libraries, New York, NY, 2023 [Online]. Available: https://library.nyu.edu/about/policies/open-metadata-policy

FIND THE MISSING PIECE:

Adding Digital Preservation to the NFT Trading Ecosystem

Pengyin Shan

University of Illinois Urbana-Champaign
USA
pengyin.shan@outlook.com
0009-0009-6309-380X

Abstract – The NFT (Non-fungible token) market is experiencing explosive growth. While artists, collectors, and crypto enthusiasts are jumping into this ecosystem, traditional collectors have found it more challenging to evaluate asset value in this NFT market than in the established collectible market. This paper navigates the reason for this problem by examining the design of NFT using ERC-721 and ERC-1155 standards, then illustrates that NFT's infrastructure makes its evaluation more challenging based on its unstable or lacking connection to the underlying digital assets, which makes the evaluation of NFT inconsistent. This paper will propose a revised business model with a workflow to add digital preservation to the NFT trading ecosystem. The paper suggests adding digital preservation clients to the Ethereum blockchain and building communication with the back-end digital preservation system to guarantee the stability of NFT's digital assets. In the end, this paper will discuss the benefits of the new business model and workflow, with potential future challenges and opportunities to the revised NFT market.

Keywords – non-fungible tokens (NFTs), digital preservation, metadata standard, NFT trading, NFT flaw

Conference Topics – Sustainability: Real and Imagined, Immersive Information

I. INTRODUCTION

NFT (Non-fungible token) has gained substantial attention growth in the past few years. The global number of users was around 36.12m in 2021 and is projected to reach 64.45m users by 2027. The worldwide revenue is expected to reach $8068.99m in 2027 [1]. OpenSea, one of the most popular NFT marketspaces, had $467,608.18 in monthly sales in Feb 2023 [2].

While the interest in NFT collections and exchanges is growing, some museums have started experimenting with preserving NFTs. 'CryptoPunk 5293' (the work's title), an NFT used to rack up about $800 million worth of sales on the Ethereum blockchain exchange, was acquired as a gift from a trustee by ICA Miami in 2021 [3]. However, the broad market has not fully recognized the necessity of digital preservation, nor has it been explored by most digital preservation organizations.

Digital preservation, which combines policies, strategies, and actions to ensure access to reformatted and born-digital content over time [4], perfectly fits the needs of NFT traders who seek stabled long-term asset value in collectibles. On the tech side, the Digital preservation systems, such as LOCKSS Software and Archivematica, have been adopted by many large-scale organizations [5]. Many academic institutions also have high-proficiency digital preservation workforces to develop and maintain these systems [6]. Besides safeguarding the digital asset in NFT, digital preservation adds accountability to the NFT and ensures the cultural heritage and financial continuity of the NFT trading ecosystem. This paper aims to analyze the reasons for some dysfunctional and deficient trading behaviors in the current NFT market, then proposes a business model and the corresponding workflow to fix the problem. The goal is to introduce digital preservation into the NFT trading ecosystem to assume, validate, and maintain the underlying digital asset of NFTs. Adopting the

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

proposed business model and workflow will provide the missing piece that the current NFT market needs to include in the traditional, thoroughly tested collectible market, and make the NFT market a more sustainable, resilient system with lasting health and vitality.

II. History Of NFT

Non-fungible tokens, or NFTs, are assets that have been tokenized on a blockchain and include unique identification metadata proving this token is distinctive. Unlike interchangeable tokens, NFTs could look identical but are non-interchangeable or non-fungible.

The history of NFT originated in 2012 when M. Rosenfeld introduced the concept of 'Colored Coin,' which represents physical assets such as money, real estate, or vehicles [7]. Two years later, Vitalik Buterin (the creator of Ethereum) and Fabian Vogelsteller introduced the ERC-20 standard for anyone to create tokens like 'Colored Coin' on the Ethereum blockchain [8].

In January 2018, etherum.org introduced the ERC-721 standard [9], which allowed people to create NFTs on the Ethereum blockchain and provided practical functionalities—these functionalities included transferring tokens from one account to another, returning the current token balance of an account and getting the total supply of the token available on the network. A significant amount of NFTs today are based on this standard. Ten months later, etherum.org published the ERC-1155 standard as the 'multiple token interface' [10], while the ERC-721 standard's token ID was the single non-fungible index. The ERC-1155 Multi Token Standard also allows each token ID to represent a new configurable token type. The majority of NFTs follow these two standards in today's market.

III. Hidden Traps in the Current NFT Ecosystem

Although many artists, collectors, and traders welcome the rising NFT ecosystem, the overall trading activities are still far less sustainable and balanced than the mature, traditional collectible market. The current NFT market challenges classic collectors and new NFT investors to fully understand and reasonably trade NFTs. In September 2022, the NFT market saw volumes down 97 percent from eight months ago [11]. To revitalize the NFT trading ecosystem, researchers have identified several issues stemming from the design of NFT marketplaces. These include a complex user interface, a lack of comprehensive documentation and guidance, and a lack of continuity between different marketplace websites [12]. However, many NFT critics still claim that NFT is just a worthless concept [13]. Their arguments highlight two major flaws that originate from the design of NFTs themselves and cannot be rectified solely by enhancing user experiences:

A. The 'Real' Digital Object Doesn't Need to Exist

A common misleading concept about NFTs is that the NFT must represent some 'visible' digital objects, like art, game gadgets, or virtual land in a Metaverse. This is not the case because both ERC-721 (defined by *ERC721Metadata* interface) and ERC-1155 (defined by *ERC1155Medata_URI* interface) make the URI to digital asset optional [9], [10], which means that NFT is not guaranteed to have a visible digital object linked to it. Some NFTs use *IPFS*, a content-addressed, versioned P2P file system widely accepted as the blockchain file storage solution [14]. However, a persistent, always-valid file location is still not mandatory for NFT trading. Even *IPFS* is not required to always be responsive by these ERC-721 and ERC-1155 standards. Moreover, the URI to digital assets typically points to an off-chain host location where the NFT buyer has no control. The file server may not guarantee the long-term existence of digital support on the file server. Accordingly, some traditional collectors or NFT beginners could be surprised one day that their NFT image shows a '404 Error' instead. They may recognize later that it is acceptable for an NFT to have an 'inconsistent' actual digital asset. Still, their confidence in continuously fair trading in the NFT market could be hurt. Eventually, this situation will block the healthy circulation of NFTs in the ecosystem.

B. The Real Value of NFT Is from Metadata Collections

The fact that underlying digital assets are not necessarily tied to NFTs creates a distinction between the NFT trading market and the traditional collectible market. Consequently, the pricing and value recognition in the NFT market differs from those in the traditional collectible market. Unlike traditional collectors who receive both proofs of ownership and physical collectibles through trading, NFT traders must understand that an NFT only

represents proof of ownership and a record of the trading experience, thus the monetary value of NFTs tends to be associated with intangible attributes derived from the NFTs' metadata. For instance, a high trading price and numerous dramatic trading stories related to an NFT may reflect the current owner's social status, power, wealth, and fame, irrespective of the existence of the digital object. As a result, NFT evaluation becomes unpredictable, making the NFT ecosystem more susceptible to asset bubbles. The uncertainty in digital asset safety and its underlying value makes NFT items hard to evaluate, hurting the overall NFT system's stabilization and long-term sustainability.

IV. Adding Digital Preservation to the NFT Trading Flow

To provide NFT owners with access to underlying digital assets, off-chain storage solutions are used, where digital objects are stored on file servers. Links in fields like ERC1155Metadata_URI or ERC721Metadata point to these off-chain file servers. These solutions enable efficient and decentralized storage of NFT-associated digital assets. However, blockchain-based off-chain storage also has drawbacks. It requires technical expertise for node infrastructure setup and maintenance, making it less accessible to non-technical users. Incentive systems tied to cryptocurrencies can be unstable, discouraging financial rewards. Insufficient active participants in the blockchain network increase the risk of data loss, posing a challenge to the persistence of digital assets. Thus, blockchain-based storage alone cannot guarantee long-term preservation [15].

But does this imply that traditional digital preservation solutions should directly compete with decentralized storage solutions in the NFT trading market? Not necessarily. Decentralized storage systems still offer distinct advantages to NFT traders, such as enhanced privacy and the absence of reliance on other nodes in the chain. Therefore, rather than engaging in a zero-sum game, digital preservation teams should focus on catering to a specific group of NFT traders who require continuous availability of associated digital assets. In a decentralized future, digital preservation players can explore new revenue streams by assuming roles as assurers, validators, and maintainers within the NFT trading ecosystem.

A. Use Case Example and Business Model

The evolving roles of digital preservation teams will provide team members with a wider market to showcase their technological strengths. One potential use case lies in the training of Large Language Models (LLMs) within the future AI market. The emergence of ChatGPT has demonstrated the significance of LLMs in achieving improved training outcomes. For instance, the initial version of ChatGPT (GPT-3) was trained on approximately 570GB of source data [16]. DALL.E with Clip, the AI system capable of generating images from text descriptions, is trained on a dataset consisting of 400 million pairs of images and text [17]. but where do these images and texts, which serve as training sources for the Large Language Models (LLMs), originally come from? Although there hasn't been a consensus on the source of the data for LLM training, web crawlers, which are tools to gather data and images from websites without guaranteeing payment to the parties being scraped, serve as another significant source of LLM training data [18].

As more companies opt to develop their own LLMs, a concerning trend emerges. Images created by artists seeking to sell their work for profit are being utilized by machine learning models without compensation. Unfortunately, there is little artists can do to prevent their creative output from being used in the production of machine-learning AI models that have the potential to replace their own work.

A new business model for NFT trading, incorporating digital preservation technologies, can help prevent the future scenario, as depicted in Fig. 1. In this ecosystem, artists would create NFTs with digital assets stored using decentralized storage solutions such as *IPFS*, while incorporating digital preservation into the workflow. The introduction of digital preservation parties in the NFT ensures the long-term guarantee of the underlying digital asset's availability. Model trainers and other data buyers, who require continuous access to the underlying data, would need to pay the NFT to access the privately encrypted digital asset for training and other profit-driven purposes. The additional value in this NFT, resulting from the involvement of the digital preservation side, is demonstrated and recognized

by artists and digital preservation teams.

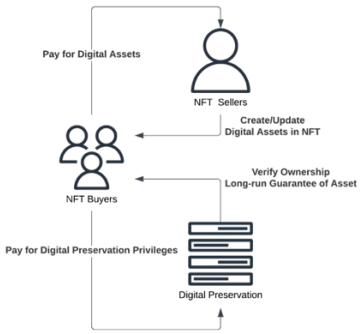


Fig. 1 NFT Trade Business Model with Digital Preservation

B. Workflow

Since blockchain has natural characteristics of immutability and decentralization, all nodes on the chain mutually watch and maintain the significant metadata of blocks. Hence a light Ethereum client should be set up on the digital preservation side to keep connecting with the Ethereum blockchain. The *Metadata JSON Schema*, part of ERC standards for NFT [9],[10], defines the digital object location and other metadata for the underlying digital asset. The digital preservation system should thus store this part of metadata to ensure the proper amount of data is kept off-chain instead of adding a heavy burden to the existing storage infrastructure.

The proposed phases of the new trading flow are stated by 1) NFT's creation and updating and then 2) an NFT's owner requesting verification of visible digital assets in NFT, as illustrated in *Fig. 2:*

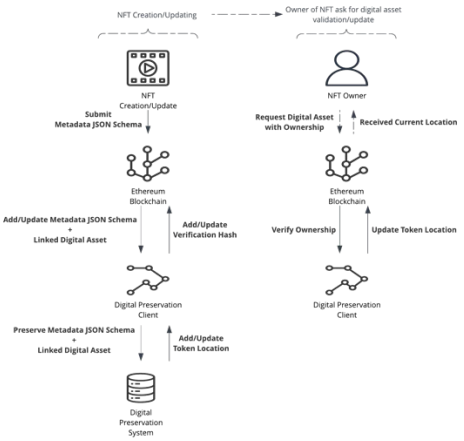


Fig. 2 NFT Trade Flow with Digital Preservation

3. A. NFT Creation or Updating

In this stage, the NFT is created or updated. An example NFT is a game card with a football player's score in alignment with the ERC-721 standard. When the card is made, the football player's initial score is shown on the GIF of the card. The GIF is stored as the image attribute, retrieved from the *tokenURI (unit256 _tokenId)* function under the ERC721MetaData interface. The image attribute will point to the updated GIF with a new score if the football player's score is updated. In these scenarios, the NFT's change will be admitted and recorded by the Ethereum blockchain. As a blockchain node, the digital preservation client will automatically be notified. The client then informs the data preservation system to update and store the updated digital asset. After the digital asset is well preserved, the digital preservation client submits an encrypted message to the blockchain to signify that the digital asset has been well maintained, including the preservation client's public address. All nodes record this verification to avoid future debuts. The digital preservation team is the digital asset watcher and keeper in this process.

B. The Owner of NFT Asks for Digital Asset Validation

This scenario occurs when the NFT owner notices that the *Metadata JSON Schema* does not reflect the correct digital asset. For example, the collector/owner of the game card sees that the GIF does not exist anymore. Then, the NFT owner messages the digital preservation address to ask for verification. The digital preservation client confirms the ownership using the NFT's metadata, then updates the digital asset address and sends the verification message again. This process is automatically broadcasted to blockchain nodes. By doing this, the underlying digital assets are successfully ensured in the NFTs, avoiding the value's significate devaluation. The NFT owner and potential buyers can continue to assess the value of NFT using the certified, industrial-trusted, and regulated guidelines. This flow will increase the number of healthy trades and improve NFT market quality sustainably.

5. C. Benefits of The New Business Model and Workflow

The business model and workflow proposed in this paper fulfill the needs of all parties involved in the NFT trading process. Artists and NFT creators receive financial rewards for their creativity while safeguarding their intellectual property, incentivizing them to contribute more original ideas to the art market. Conversely, NFT buyers, who require long-

term assurance of the underlying digital asset, can access mature services provided by the digital preservation team. Additionally, the digital preservation team not only benefits from new revenue streams in the business model but also has opportunities to become part of the broad NFT community and gradually establish a reputation in the Web 3.0 era.

Meanwhile, this designated flow will not disrupt the current NFT ecosystem because it does not add anything new to the current NFT standards, such as ERC-721 and ERC-1155. Instead, this flow will add digital preservation as a distinct, trusted resource for interested NFT creators/owners to seek verification and safe, long-term digital asset maintenance.

The digital preservation client joins the blockchain network, meaning that if the blockchain upgrades in the future, the digital preservation clients can participate in the upgrading decisions, have chances to vote as other nodes and enjoy the benefits of the new blockchain infrastructure after upgrading.

V. Potential Challenges And Future Works

The NFT market is still in the early stages of development. The proposed trading business model and workflow could face many potential challenges brought mainly by the uncertainty of the overall blockchain ecosystem. Two significant obstacles include regulatory uncertainty and possible environmental criticism.

NFT has yet to be entirely accepted in many countries and faces legal side ambiguities. For example, Russia does not support NFT in any form [19]. Digital preservation parties should prepare for possible debates before joining the workflow, especially arguments about intellectual property.

Blockchain society is working towards a more sustainable, eco-friendly architecture. Ethereum, as the major player in the NFT market, is expected to drop by a massive 99.988% and its carbon emissions by 99.982% after it switched to the 'Proof of Stake' algorithm in later 2022 [20]. However, governments and NGOs are still concerned about the massive energy consumption supporting blockchain systems' computational power [21]. The data preservation team should continuously collaborate with other stakeholders to optimize the workflow to become more energy efficient.

Besides seeking solutions to these two challenges, the future work for this paper can also focus on technical implementation details, such as selecting the appropriate preservation technology for different NFT digital assets. Technologies like archival packages, emulations, or media-independent assessments have been proposed for NFT digital preservation [22]. It would be valuable to verify their applications in various scenarios through further research.

Additionally, special use cases within the new business model should be carefully considered and evaluated when designing the technical system. For example, if a blockchain becomes compromised or inactive due to security attacks, legal requests, or lack of participants, NFT owners still require continuous access to the underlying digital assets. Should the digital preservation team maintain a full local copy of the blockchain to prepare for such cases? Alternatively, should they develop a cross-chain solution to back up NFTs on another active chain? Will the cross-chain solution break the creators' original purpose in the art [23]? There is ample opportunity for future researchers to explore and navigate the technical implementations within the framework of the new business model.

VI. Conclusion

Compared to the traditional healthy collectible market, the NFT ecosystem lacks long-term stability to the NFT's off-chain underlying digital asset, which can lead to irrational trading and market phishing. Digital preservation will make up this missing piece by ensuring that digital support remains accessible, usable, and trustworthy over time. The designated business model and workflow in this paper add data preservation teams to the NFT ecosystem. Data preservation will actively guide, verify, and steward digital assets to guarantee NFT owners get the real value they seek from the NFT item. By adding the critical digital preservation piece into the ecosystem, the NFT market will keep a healthy trading environment in the long run.

1. REFERENCES

[1] NFT - Worldwide, n.d. [Online]. Available: https://www.statista.com/outlook/dmo/fintech/digital-assets/nft/worldwide

[2] Largest NFT marketplaces based on all-time sales volume in the previous 30 days as of February 22, 2023 (in Ethereum and U.S. dollars) [Graph], NFTGo, February 22, 2023. [Online]. Available: https://www.statista.com/statistics/1274843/nft-marketplaces-with-highest-volume/?locale=en

[3] F. Nayeri, "NFTs, on the Decline Elsewhere, Are Embraced by Some Museums," *The New York Times*, Nov. 30, 2022. Accessed: Mar. 07, 2023. [Online]. Available: https://www.nytimes.com/2022/11/30/arts/design/nfts-museums.html

[4] ALA Annual Conference, "Definitions of Digital Preservation," *Association for Library Collections & Technical Services (ALCTS)*, Feb. 21, 2008. https://www.ala.org/alcts/resources/preserv/defdigpres0408#:~:text=Medium%20Definition (accessed Mar. 08, 2023).

[5] V. Reich and D. S. H. Rosenthal, "LOCKSS (lots of copies keep stuff safe)," *New Review of Academic Librarianship*, vol. 6, no. 1, pp. 155–161, Jan. 2000, doi: https://doi.org/10.1080/13614530009516806.

[6] H. Hockx-Yu, "Digital preservation in the context of institutional repositories," *Program*, vol. 40, no. 3, pp. 232–243, Jul. 2006, doi: https://doi.org/10.1108/00330330610681312.

[7] M. Rosenfeld, "Overview of Colored Coins," 2012. Available: https://bitcoil.co.il/BitcoinX.pdf

[8] F. Vogelsteller and V. Buterin, "EIP 20: ERC-20 Token Standard," *Ethereum Improvement Proposals*, Nov. 19, 2015. https://eips.ethereum.org/EIPS/eip-20

[9] W. Entriken, D. Shirley, J. Evans, and N. Sachs, "EIP 721: ERC-721 Non-Fungible Token Standard," *Ethereum Improvement Proposals*, Jan. 24, 2018. https://eips.ethereum.org/EIPS/eip-721

[10] W. Radomski, A. Cooke, P. Castonguay, J. Therien, E. Binet, and R. Sandford, "EIP 1155: ERC-1155 Multi Token Standard," *Ethereum Improvement Proposals*, Jun. 17, 2018. https://eips.ethereum.org/EIPS/eip-1155

[11] S. Shukla, "NFT Trading Volumes Collapse 97% From January Peak," *Bloomberg.com*, Sep. 28, 2022. Accessed: Jun. 29, 2023. [Online]. Available: https://www.bloomberg.com/news/articles/2022-09-28/nft-volumes-tumble-97-from-2022-highs-as-frenzy-fades-chart#xj4y7vzkg

[12] S. V. Murphy Caxton, K. Naveen, R. Karthik, and S. S. Bama, "User-Centered Evaluation and Design Suggestions for NFT Marketplaces," *IEEE Xplore*, Jul. 01, 2022. https://ieeexplore.ieee.org/abstract/document/9850815 (accessed Mar. 09, 2023).

[13] E. Kraizberg, "Non-fungible tokens: a bubble or the end of an era of intellectual property rights," *Financial Innovation*, vol. 9, no. 1, Jan. 2023, doi: https://doi.org/10.1186/s40854-022-00428-4.

[14] J. Benet, "IPFS - Content Addressed, Versioned, P2P File System," *arXiv.org*, 2014. https://arxiv.org/abs/1407.3561

[15] C. Dupres, "IPFS, Filecoin and the Long-Term Risks of Storing NFTs," *www.coindesk.com*, Jan. 20, 2022. https://www.coindesk.com/layer2/2022/01/20/ipfs-filecoin-and-the-long-term-risks-of-storing-nfts/

[16] A. Iyer, "Behind ChatGPT's Wisdom: 300 Bn Words, 570 GB Data," *Analytics India Magazine*, Dec. 15, 2022. https://analyticsindiamag.com/behind-chatgpts-wisdom-300-bn-words-570-gb-data/ (accessed Jun. 29, 2023).

[17] K. Johnson, "OpenAI debuts DALL-E for generating images from text," *VentureBeat*, Jan. 05, 2021. https://venturebeat.com/business/openai-debuts-dall-e-for-generating-images-from-text/

[18] A. Tzeveleka, B. Pistone, O. Cruchant, and G. Nachum, "Training large language models on Amazon SageMaker: Best practices | AWS Machine Learning Blog," *aws.amazon.com*, Mar. 06, 2023. https://aws.amazon.com/blogs/machine-learning/training-large-language-models-on-amazon-sagemaker-best-practices/#:~:text=LLM%20developers%20train%20their%20models (accessed Jun. 19, 2023).

[19] L. V. Astakhova and N. V. Kalyazin, "Non-Fungible Tokens (NFT) as a Means and Object of Ensuring Information Security," *Automatic Documentation and Mathematical Linguistics*, vol. 56, no. 3, pp. 116–121, Jun. 2022, doi: https://doi.org/10.3103/s0005105522030062.

[20] Crypto Carbon Ratings Institute, "Sep REPORT The Merge Implications on the Electricity Consumption and Carbon Footprint of the Ethereum Network," https://carbon-ratings.com, Sep. 2022.

[21] F. Valeonti, A. Bikakis, M. Terras, C. Speed, A. Hudson-Smith, and K. Chalkias, "Crypto Collectibles, Museum Funding, and OpenGLAM: Challenges, Opportunities and the Potential of Non-Fungible Tokens (NFTs)," *Applied Sciences*, vol. 11, no. 21, p. 9931, Oct. 2021, doi: https://doi.org/10.3390/app11219931.

[22] J. Bell, R. Harsanyi, and J. Ippolito, "RIGHT CLICK TO PRESERVE Preservation, NFTs, and Distributed Ledgers," Oct. 2022. Accessed: Jun. 29, 2023. [Online]. Available: https://osf.io/zt6ex/

[23] J. Ippolito, "Crypto-Preservation and the Ghost of Andy Warhol," *Arts*, vol. 11, no. 2, p. 47, Mar. 2022, doi: https://doi.org/10.3390/arts11020047.

THE REMATRIATION PROJECT

Building Capacity for Community Digital Archiving in Northwest Alaska

Erin Yunes
Virginia Tech
United States
eyunes@vt.edu

Cana Uluak Itchuaqiyaq
Virginia Tech
United States
cana@vt.edu

Kara Long
Virginia Tech
United States
karal@vt.edu

Abstract – Directed by an Inuit-led and serving tribal organization, Aqqaluk Trust, in the frontline hub-community of Kotzebue, Alaska, the Rematriation Project: Restoring and Sharing Inuit Knowledges aims to create capacity for and access to digital archives related to Inuit cultural, tribal, and scientific knowledges and history to assist tribes and communities in developing localized, culturally appropriate approaches and solutions to their needs. In partnership with a team of scholars from Virginia Tech (itself led by an Iñupiaq scholar from Kotzebue)—the goal of this project is to empower Indigenous communities through the lens of Indigenous data and research sovereignty to collect, control, interpret, and benefit from data that originates from their communities. The Rematriation Project operates on a foundation of community-first, community-led decision making that emphasizes Indigenous Data Sovereignty practices. This paper outlines the goals and initiatives of the first phases of the project.

Keywords – Community Archives, Preservation, Indigenous Data Sovereignty, Capacity Building, Equitable Research

Conference Topics – Digital Accessibility, Inclusion, and Diversity; Sustainability: Real and Imagined

I. INTRODUCTION

As Cree-Métis scholar and librarian, Jessie Loyer, expresses in *The Collector and the Collected*, "Who has the authority to own and manage collections? … Who is granted the credibility to disseminate this information? … Indigenous communities have too often had restricted access to the information created about them and have largely been absent from the process of dissemination of these knowledges [1]. Scholars of research data stewardship and digital preservation have acknowledged the gaps between Indigenous data sovereignty and best practices for open and accessible research data with the development of the CARE Principles for Indigenous Data Governance [2]. Extractive data practices and research methods have multifaceted impacts on communities and perpetuate colonialist and inequitable power differentials [3], [4].

Climate scientists gather and preserve massive amounts of data each year from Arctic Indigenous lands. This data powers the dominant research data lifecycle model. Although several variations of the data lifecycle exist, the basic scaffolding: acquire, process, analyze, archive, disseminate, and reuse/delete, have largely become the standard practice of institutional researchers [5], [6]. The Research Data Lifecycle supports the needs of researchers and excludes Indigenous communities from exercising data sovereignty—to have ownership over the data; to consent and control how it is used; determine who has, or can have, access to it; and decide how, where, and for how long that data will be stored [1], [7], [8].

Given this research landscape, what does it mean to be a steward of community memories and archives in 2023, especially during the climate crisis? For the partners of the Rematriation Project, it means to respectfully and equitably help Indigenous communities access, engage, and preserve cultural knowledge to fulfill their self-determined needs and goals, such as accessing and consulting traditional knowledges to determine culturally-appropriate responses to climate change.

II. THE REMATRIATION PROJECT

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

Directed by an Inuit-led and serving tribal organization, Aqqaluk Trust, in the frontline hub-community of Kotzebue, Alaska, the Rematriation Project: Restoring and Sharing Inuit Knowledges aims to create capacity for and access to digital archives related to Inuit cultural, tribal, and scientific knowledges and history to assist tribes and communities in developing localized, culturally appropriate approaches and solutions to their needs. In partnership with a team of scholars from the Virginia Tech Department of English and University Libraries (itself led by an Iñupiaq scholar from Kotzebue)—the goal of this project is to empower Indigenous communities through the lens of Indigenous data and research sovereignty to collect, control, interpret, and benefit from data that originates from their communities. This project began from a series of informal conversations about community needs in relation to community experiences with academic research between researchers and Aqqaluk Trust staff [9]. Through these conversations, it was determined that community digital archiving needs existed in the region, and that the process and skills related to digital archiving complemented other community goals and needs [9].

Kotzebue is a rural Iñupiat coastal community located above the Arctic Circle that serves as a central location for ten surrounding villages. This region is currently facing the devastating effects of rapidly accelerating climate change. Tribal communities are encountering more frequent destructive storms, fire, and flooding, putting them and their tribal histories and land stewardship at great risk. It is imperative to create accessible, digital versions of valuable and threatened knowledges. Creating digital archives is one part of the solution, developing local capacities for digital archiving is another.

The Rematriation Project fulfills these needs for the Iñupiat of NW Alaska and also provides a transferable model and materials for other communities to use for their own self-determined needs. In order to accomplish project objectives, partners center the needs and values of the community so that community members are able to make informed decisions about:

- with whom and how to share their knowledge,
- the consequences and impacts of making Inuit knowledge interoperable with other dataset and collections in a digital environment, and
- culturally appropriate and meaningful arrangement and description.

III. Centering Community

Alaska Native (Unangax) scholar, Dr. Eve Tuck, describes Rematriation as "… concerned with the redistribution of power, knowledge, and the dismantling of settler colonialism" [10]. Rematriation encompasses Indigenous-led methods of data sovereignty as well as restoring (and sharing) cultural knowledges back with Indigenous peoples. As this methodology grounds the project, partners and team members work under a community-first approach, which focuses on building trusting relationships and partnerships within the team and community, centers community-needs and values, and implements community-led/advised decision making.

The first step in the project team's methodology has been to engage Cultural Humility as a framework [11], prior to beginning any research or activities. Cultural Humility is the continuous process of reflection and self-evaluation of your history, background and objectives; a committed renewal to learning from the community, challenging your own biases and beliefs, and restoring imbalances; and holding yourself accountable in understanding how your history, biases, and beliefs influence your actions and impact the community [12], [13]. Although much of this work is an internal process, the project team operationalizes the Cultural Humility framework through open dialog and discussions in weekly check-in meetings. These meetings not only give the team the opportunity to discuss upcoming initiatives but understand how this work, and our roles within the work, benefits the community and supports its values and self-determined needs. This approach leads to strengthened relationships within the team prior to visiting communities for "official" research activities, and included an informal, relationship-building community visit in April 2023. The weekly check-in meetings are also instrumental in conceptualizing a community-led and developed digital archive and culturally appropriate archival curriculum. Part of these meetings is dedicated to a discussion of Iñupiat

Ilitqusiat, Inupiat cultural values, and how they can be foregrounded in the Rematriation Project's work.

IV. Rematriation Project Goals

As the NW Arctic region faces the devastating effects of rapidly accelerating climate change, it is critical that communities not only have the resources to preserve their knowledge but retain, recover, and utilize the data collected by academic institutions. The Rematriation Project has developed a multi-phase process to accomplish its goals in creating capacity for and access to digital archives related to Inuit culture and knowledges:

A. Digitize tribal materials from Kotzebue to create a scalable model for community digital archiving.

With the help of the digitization lab at Virginia Tech Libraries, the team has digitized a small collection of papers and other artifacts of the deceased Siberian Yupik leader, Caleb Pungowiyi. These materials were donated to this project by the Pungowiyi family in Kotzebue and Caleb Scholars Program and are currently located at Virginia Tech. After the digitization is complete, the Pungowiyi materials and corresponding data will be returned to his family in Kotzebue, following a post-custodial and collaborative model for community archives. This collection provides the team with strong examples for our work that can be scaled by Aqqaluk Trust and others to meet future and broader rematriation needs.

Caleb Pungowiyi worked tirelessly to have Indigenous perspectives, needs, and knowledges included as part of major policy discussions about climate change and other conservation issues. His impact on policy discussion included the US Marine Mammal Commission and the Arctic Council, and his voice still resonates across Arctic advocacy and research circles. Pungowiyi's materials are a rich source of information about climate change and Indigenous methods of recognizing and adapting to climate change. They contain specific scientific knowledge of Inuit homelands and its changes over time that complement and extend western science. Access to these materials is culturally powerful, scientifically significant, and of critical importance to the future of the region. The Pungowiyi collection provides a strong model of the types of materials and knowledge resources that exist in Inuit communities—materials that need to be digitally preserved and accessible.

B. Increase community capacities in digital archiving and data literacies through the creation of guides and curriculum, including cataloging metadata and using existing online archival tools.

The project team is developing a series of storyboard scripts and user personas and scenarios to construct a culturally-appropriate curriculum for creating, contributing to, and using community digital archives. As with our approach to the Pungowiyi collection, the team is seeking to incorporate concepts and frameworks from educational leadership, archives and cultural heritage sectors, and the growing body of literature on cross-functional teams and team-based collaborations in the Arctic.

Culturally appropriate instruction in community archiving must address intersecting interests and needs. For example, the interests of the project team in sharing our work with other researchers and archives practitioners and the needs and rights of the community to protect culturally sensitive traditional knowledges. Descriptive and structural metadata must be meaningful to the community, as it enables users to navigate and make sense of the digital collection. Co-creating both a workflow for and standards to guide metadata creation and management that is appropriate for the community is a more complex process than selecting a metadata standard to adopt and implement whole-scale. As the team has discussed and developed metadata work, we have had to adopt a system agnostic, principles-based approach. We are continuously interrogating metadata tools and practices that have wide adoption in libraries and archives but are not necessarily in alignment with community needs or goals. As the community begins to digitize their own materials, it will help in communicating the important cultural, contextual, and historical information about each object. It is important to emphasize that the Rematriation Project was established with a strong commitment to Indigenous Data Sovereignty. Unlike other models for post-custodial archives, any data generated throughout or after the project, including digitized materials, will adhere to the principles of CARE and OCAP [2, 8], meaning that the community has ownership, control,

and possession of the digitized materials and data. The pilot collection was digitized at the University Libraries to serve as a teaching and training model, and the community will determine who can have access to the information generated from the digitization and description of the materials. Further, the protection of traditional and Indigenous Knowledges in relation to data and "Intellectual Property" has been formalized in legal contracts between Virginia Tech and Aqqaluk Trust.

In order to develop the curriculum to support the community throughout the digitization and archival process, the team will use the "I do, we do, you do" differentiated learning method: *I do*, demonstrate the digitization of a model post-custodial collection, the Pungowiyi papers; *We do*, work together to develop community workshops to identify and prepare materials for supporting the digitization of community archives; and *you do*, provide support while pilot participants work with their own materials and apply their learning toward their development as trainers [14]. This model is also consistent with Inuit cultural practices of education and experiential learning.

This is a method of instruction that can be especially useful in establishing relationships with community members that can evolve over time—from introducing a concept, a model, and an example; working together towards a common goal of contributing to a community archive; and providing support into the future as participants gain the experience and confidence to initiate and sustain projects with or without partners.

C. *Use community digital archives to design and test an online library (i.e., a website that hosts local digitized materials, provides access to existing archives, and can track new research requests) that is specifically created for Inuit users to access community databases and connects outside researchers to community liaisons.*

Although the process of designing and testing a NW Arctic Cultural Digital Library falls in the third phase of the project, currently the team is conducting a landscape analysis of websites and digital archives with similar missions to understand what features organizations are using, on what platforms these sites are built on, and how they perform in a broadband environment similar to what the communities experience in Alaska. So far, 50 digital archives have been analyzed for design and organization of content; performance on smaller mobile screens and lower bandwidth connections; and accessibility features, such as how they help visually-impaired users or not. From this exercise, the team will move into the next phase, engaging with various projects from the analysis to gain insights into user and administrator experiences. The team will ultimately create a presentation for the community that discusses build options, including costs, data storage, security features, data backup and recovery, and user experience and engagement features. The objective is to present multiple approaches for constructing the archive, with a strong focus on protecting community data and ensuring long-term data sovereignty. This is a further extension of post-custodial archival practice, in which the community partnership with the University Libraries extends beyond the digitization of a collection to also include consultation, training, and support that seeks to be responsive to community needs.

V. Conclusion

Researchers have traveled to Iñupiat lands to study the environment and culture for decades. These knowledges have not always been shared back with communities in ways that can be easily accessed, understood, or used to help with the community's self-determined needs. A vast amount of Indigenous knowledges currently live behind academic paywalls and are owned and controlled by academic institutions.

Within communities, Iñupiat traditional knowledges are documented in various ways and stored in homes, schools, and organizational buildings. However, these collections are scattered and in jeopardy of loss from housing and building insecurity, deteriorating infrastructure, mold, inadequate storage, and environmental crises. Making accessible digital versions of these valuable and threatened knowledges is imperative. The goal of the Rematriation Project is to provide Iñupiat communities with targeted, culturally appropriate capacity building that hones, develops, and complements local skills related to digital archiving and digital literacies as well as to produce a transferrable protocol for researchers that prioritizes Indigenous data and research

sovereignty, which communities can use for their own self-determined needs. The creation of the NW Arctic Cultural Digital Library will establish a platform and a space for publishers to begin the work of returning Inuit cultural knowledges back to their people.

The Rematriation Project operates on a foundation of community-first, community-led decision making that emphasizes Indigenous Data Sovereignty practices. Community workshops are a critical part of building strong, long-lasting relationships and trust. The ultimate goal of the Rematriation Project is to build capacity for community digital archiving in Kotzebue and to the surrounding villages. The digitization and creation of the Pungowiyi digital collection has been our way of working towards this goal. The pilot collection also gives the project team and community members a benchmark for scaling future digital projects. Part of the process of building and participating in the workshops is to support community participants in developing and exercising digital literacy skills that can be transferable to future projects and support the community in making self-directed decisions about data stewardship.

1. REFERENCES

[1] J. Loyer, "Collections are our relatives: Disrupting the singular, white man's joy that shaped our collections," in *The Collector and the Collected: Decolonizing Area Studies Librarianship*, Sacramento, CA: Library Juice Press, 2021.

[2] S. R. Carroll *et al*., "The CARE Principles for Indigenous Data Governance," *Data Science Journal*, vol. 19, no. 1, Art. no. 1, Nov. 2020, doi: 10.5334/dsj-2020-043.

[3] J. Sadowski, "When data is capital: Datafication, accumulation, and extraction," *Big Data & Society*, vol. 6, no. 1, 2019, doi: 10.1177/2053951718820549.

[4] T. Kukutai and D. Cormack, "'Pushing the space': Data sovereignty and self-determination in Aotearoa NZ," in *Indigenous Data Sovereignty and Policy*, Routledge, 2020.

[5] "Research Lifecycle Guide," Princeton Research Data Service. https://researchdata.princeton.edu/research-lifecycle-guide/research-lifecycle-guide (accessed Mar. 09, 2023).

[6] "8 Steps in the Data Life Cycle | HBS Online," *Business Insights Blog*, Feb. 02, 2021. https://online.hbs.edu/blog/post/data-life-cycle (accessed Mar. 09, 2023).

[7] J. Taylor, T. Kukutai, and Australian National University. Centre for Aboriginal Economic Policy Research, "Data sovereignty for Indigenous peoples: Current practice and future needs," in *Indigenous data sovereignty: toward an agenda*, 1 online resource (xxiii, 318 pages): illustrations vols., Acton, ACT, Australia: Australian National University Press, 2016, pp. 1–22. Accessed: Mar. 09, 2023. [Online]. Available: https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=3093595

[8] "The First Nations Principles of OCAP," *The First Nations Information Governance Centre*. https://fnigc.ca/ocap-training/ (accessed Aug. 26, 2022).

[9] C. Q. Kramer and C. U. Itchuaqiyaq, "Participatory Research & Indigenous Leadership—A Model for Equitable Arctic Research from Kotzebue," presented at the IARPC Collaborations, May 18, 2023. Accessed: Jun. 12, 2023. [Online]. Available: https://www.youtube.com/watch?v=rLWXWbDlpIk

[10] E. Tuck, "Rematriating Curriculum Studies," *Journal of Curriculum and Pedagogy*, vol. 8, no. 1, pp. 34–37, Jan. 2011, doi: 10.1080/15505170.2011.572521.

[11] C. U. Itchuaqiyaq, C. Lindgren, and C. Q. Kramer, "Decolonizing community-engaged research: Designing CER with cultural humility as a foundational value," *Communication Design Quarterly*, n.d.

[12] N. Andrews, S. Kim, and J. Watanabe, "Cultural Humility as a Transformative Framework for Librarians, Tutors, and Youth Volunteers: Applying a lens of cultural responsiveness in training library staff and volunteers," *Young Adult Library Services*, vol. 16, no. 2, pp. 19–22, Winter 2018.

[13] S. Y. Leung and J. R. López-McKnight, *Knowledge justice: disrupting library and information studies through critical race theory*. Cambridge, Massachusetts: The MIT Press, 2021.

[14] L. Bowgren and K. Sever, *Differentiated Professional Development in a Professional Learning Community*, Reprint edition. Bloomington, IN: Solution Tree, 2009.

THE LONG AND WINDING ROAD

Implementing an Open-Source Workflow

Karyn Williamson
abrdn plc archive
Scotland
Karyn.Williamson@abrdn.com
0000-0003-0426-2406

This paper will discuss the recent work of the abrdn archive to create and put in place a fully open-source workflow, the barriers faced, and what was learnt from the experience. abrdn plc, an Edinburgh based investment company, is the first financial institution to implement a fully open-source digital preservation workflow. The following text provides an analysis of the work undertaken by the archive thus far, and argues that the least important aspect of implementing an open-source workflow is the software. Our journey has definitely been a long and winding road with many wrong turns, misleading directions, and occasional problems with running out of fuel, but the outcome has been positive and archive colleagues are keen to find out where the road will lead them next.

Keywords – open-source, workflow, software, analysis

Conference Topics – We're all in this together; From theory to practice.

I. INTRODUCTION

abrdn plc, is an Edinburgh based investment company, which currently manages over 550 billion in assets with over 800 investment managers spread over 30 locations globally. The company was formed when Standard Life plc merged with Aberdeen Asset Management plc in 2017 to form Standard Life Aberdeen. A rebrand and name change to abrdn took place in 2021.The abrdn archive is responsible for the records of all these companies and their subsidiaries. The archive contains over 163 cubic meters of physical records and approximately 2tb of digital material, including both born digital and digitized.

For the last 3 years the archive has worked with a range of internal stakeholders to put in place a robust, scalable fully open source digital preservation workflow which would link in with the procedures already established for physical collections. This paper will document and analyze the steps taken on this journey, the outcome as it currently stands and plans for the future. This paper will serve as a case study for those interesting in implementing an open-source solution within their own organizations', and will highlight the fact that with careful and focused planning, any institution can put an open-source workflow in place to protect their digital holdings.

II. CONTEXT OF OPERATION

The context in which the archive operates is essential to understanding the workflow put in place and the reasons why the decisions described hereafter were made.

Since abrdn was formed, it has been working through a landscape changing company transformation, including a full rebrand project which included a change of name. This has led to the work of the archive taking lower priority in terms of overall company strategy. This has had a direct impact on the budget and resource available to the archive for digital preservation work; both of which were already in short supply.

Although abrdn is a global company, there is no global archive provision and colleague resource has been limited to no more than 2 full time archivists at any one time since the archive function was formed. Less than 1/3 of this resource is allocated to digital preservation work, which had resulted in a sporadic at best approach to digital preservation. The approach at this time was to collect the digital records documented in our collecting policy, but not

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

to process them in any way. This led to a backlog of digital processing work, for which the archive team had limited resource available and little knowledge that could be used to processing these records. A business case for digital preservation system funding was put forward which detailed the gold (preservica), silver (archivematica) and bronze (fully open source) routes to implementing a digital preservation workflow. The business case levels were defined based on the resource required to implement each option, ease of use for staff, the training and knowledge available about each option and the number of records that could be processed per day using each option. Ultimately, the decision was made to implement the bronze option with no other resource approved other than that needed for membership of the Digital Preservation Coalition as. It was recognized that the current skillset of the archive was not in a place to carry out this work without guidance and help. This only became more apparent as the project progressed, particularly in terms of security restrictions.

The cyber security restrictions in place to keep abrdn data safe are among the highest in the world and until this work began, the company had actually issued a blanket ban on open-source software being used on their systems. These restrictions combined with the financial regulations and compliance needs of the company were the backdrop to an already complex problem and it was clear that extensive research and planning would be needed at the outset in order to put a successful workflow in place.

III. Workflow planning

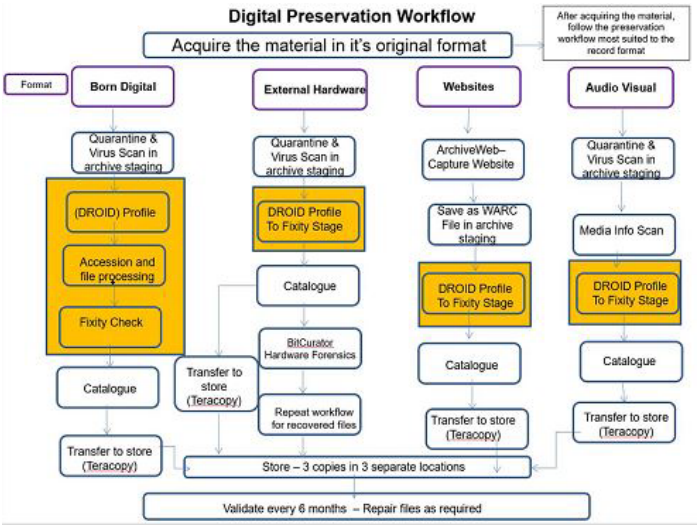
Initial approaches to working out an open-source workflow focused mainly on software, and it quickly became clear that this was the wrong place to start. Focusing on how tasks would be done meant we very quickly lost sight of why we were doing the task in the first place. Despite being completely new to the world of digital preservation, we had tried to start at one of the most technical points in the process and our lack of technical knowledge led to us becoming overwhelmed and confused. Taking a step back and beginning with a holistic view of what was required allowed us to start at the beginning and work through what would be required at each phase of the workflow. This step-by-step approach also allowed for an ongoing period of research which help build up the skillsets and technical knowledge that would be required later in the project. Working in this way also made it clear that despite being at the beginning of our digital preservation journey, some of the processes and systems already in place across the company were compatible with our requirements. Although purely coincidence, this gave team morale a boost and helped us move forward with the process with more confidence and purpose. Our confidence was increased further when we started using the resources provided by the Digital Preservation Coalition.

Many of the resources freely provided by the Digital Preservation Coalition are incredibly useful for beginners to the sector who are looking to implement a system in their organization.[1] The DPC RAM assessment helped us to map out our current position, and what would be needed to get us to where we wanted to go, and the DPC handbook outlined the steps we would need to take to get there. Consulting with the company IT department was also beneficial as they helped us understand where our workflow could merge with current company procedure and where a new process would need to be developed. Having access to this range of technical knowledge combined with input from the digital preservation community was essential to the success of the program. To build up the knowledge of our current position further, we combined the results of the DPC RAM assessment with the results of an NDSA levels of Digital Preservation assessment.[2] Completion of the NDSA levels matrix gave us a basic overview of where we were technically and how the requirements of the workflow being implemented might impact on the wider company. Having this knowledge would be essential when we reached the next phase and began looking at software options.

Once the planning phase was complete, Fig,1 below is the workflow we chose to implement. The orange section highlights the repetition of the full workflow detailed in the born digital flow in the smaller orange section in the other flow lines.

[1] Mentioned resources plus many more available here, Digital Preservation - Digital Preservation Coalition (dpconline.org)

[2] NDSA guidance and matrix available here, Levels of Digital Preservation (ndsa.org)



At this point, no software was included on the workflow and so work began on finding the programs that were best suited to the tasks we wanted to perform.

As mentioned previously, software can be one of the most confusing and intimidating parts of the digital preservation process. The sheer volume of programs available is daunting in itself. In addition, technical experience varies greatly from one digital preservation practitioner to the next and not every open-source tool available is beginner friendly. The main digital preservation practitioner on this project trained firstly as an archivist working with physical papers and working with digital records was a learning on the job process which slowed things down significantly. The archive entered a prolonged period of testing to work out what programs best suited each role and finally settled on the software documented in Fig.1. The security restrictions and regulations the archive operates under made this section of the work plan particularly problematic. Although the company had approved the use of open-source software by this stage, it was under the caveats that any program being put on our systems had to be approved by the software approval board (SAB) and they would only approve programs which had been robustly tested and were being used as part of the final workflow. The SAB would only approve software with a graphic user interface as security restrictions meant no access to programs running via the command line. To get around this I was given a laptop with no connection to the company servers with various data sets loaded onto it that I could use to test any program I wanted to and then confirm intended use with the SAB so it could go through the approval process.

This stage of the process took over 15 months to complete. The learning process around each program tested was complex and long due to the knowledge base of the archivist at the beginning of the project. Upon reflection, Implementing this workflow made it clear that time is not always a good indicator of the success of a project and that taking the time to get the right result for the archive was more important than having a set-up complete within a set timeframe. This section of the journey was particularly bumpy with a few pit stops taken to re fuel, re group and look at things from a different perspective to find the answers we needed.

IV. Workflow Testing

With all the chosen software in place, the next step was end to end workflow testing to ensure that each section worked not only independently but also as part of the workflow. Due to the earlier intensive planning and testing, this stage was relatively smooth. It was only at this point when the workflow was being tested and found to be working that the team felt like they were actually 'doing digital preservation'. From a beginner's perspective, it can seem like you need to be doing the technical parts and physically preserving records and using software to be carrying out digital preservation tasks, but this is definitely not the case. The planning and testing work carried out was essential to the success of the workflow being implemented and having worked through this process, it is clear that the earlier preparation work was more important than the final test phase. Very few issues required fixing and the workflow is now in use to archive abrdn records as part of BAU work tasks (business as usual). It took over two and a half years to get to the point where preserving the digital records of abrdn was integrated as part of the day job but looking back, the journey was worth it. We are the first financial institution to implement a fully open-source workflow (that we are aware of) and the archive as a whole now has far more knowledge about dealing with digital records and the work involved in preserving them.

Although the workflow is in place and working well, we are still on the long and winding road with the end not quite yet in sight. The process outlined in this paper works well for small digital collections,

but automation of various tasks is required to enable the archive to process larger collections. There is also work be done around the file formats identified by DROID that we can't yet open or preserve. This includes a range of files stored on DVD, CD and floppy discs. The next step after this will be to look at our digitization processes and document them in a similar workflow that can link to the current digital preservation workflow.

V. CONCLUSIONS

The story of the journey the abrdn archive has been on is an important one, because it highlights that anyone in any organization can implement a digital preservation workflow that covers all required bases. The planning phase, as this paper documents, is the most important part of the process and shouldn't be rushed. Networking and attending training events held across the digital preservation sector is also invaluable to those at the beginning of their own digital preservation story. The connections made to those with more experience and listening to what others in the field are working on helps to challenge the imposter syndrome that many beginners feel when starting on their digital preservation journey as well as helping them to connect with others working in a similar space who they can learn from and collaborate with. The resources available online, including those from the Digital Preservation Coalition, are also an invaluable step on any beginner expedition into preserving digital records. The main lesson learnt on this journey, is that software selection is a very small part of the overall work involved in creating a digital preservation workflow. In addition, the software testing phase runs far smoother when informed by the planning phase, as the practitioner will know what is required of each program and can quickly discard programs that don't meet this predefined criteria. To refer back to the title of this paper. The journey from complete beginner with very little knowledge to a full y working, continually developing digital preservation workflow for the abrdn archive has definitely been a long and winding road, and one that still has a lot of twists and turns before its final destination will be reached. But this shouldn't put off other beginners looking to start their own journey. Lau Tzu stated that "The journey of a thousand miles begins with a single step".[3] Never has a truer statement been made that describes working in digital preservation. Taking each step on the journey one at a time and making sure you have a strong foothold before moving on, and not being afraid to ask for a helping hand (or a rest!), when needed will help ensure long term success and a fully functioning open-source digital preservation that meets all requirements. The most important step on this journey, is the one that gets you started.

[3] Tzu Lao, *Tao Te Ching*, 1933

EaaSI PRESERVATION OF MOBILE APPLICATIONS

Progress with the long-term preservation of access to mobile applications using the EaaSI platform

Euan Cochrane
Yale University Library
Country
euan.cochrane@yale.edu
0000-0001-9772-9743

Jurek Oberhauser
OpenSLX
Germany
jurek@openslx.com
0000-0003-4542-7959

Rafael Gieschke
University of Freiburg
Germany
rafael.gieschke@rz.uni-freiburg.de
0000-0002-2778-4218

Abstract – Mobile devices have revolutionized computing and democratized access to it. The applications we use on our mobile devices play a critical role in shaping our online experiences, our culture, our politics, and our access to information. Mobile applications are also widely used for data gathering and asset management in many domains from scientific research to infrastructure maintenance. With such a wide-reaching impact it is critical that the preservation community is able to maintain access to mobile applications for future generations. In this short paper we outline progress in using the Emulation as a Service Infrastructure (EaaSI) platform to run obsolete versions of the Android operating system in virtualization and emulation in order to maintain access to mobile applications. We also detail the current limitations of virtualizing and emulating mobile devices and provide a list of future challenges to address as we move forwards with ensuring long-term access to this essential part of our history.

Keywords – emulation, mobile, apps, applications

Conference Topics – From Theory to Practice; Immersive Information

I. INTRODUCTION

Within the Emulation as a Service Infrastructure (EaaSI) platform users can already run some versions of the Android operating system using the existing QEMU emulator. Since Android is a variant of the well-supported Linux-based operating system family, Android versions made for the IBM PC platform using the x86(-64) architecture [1] can generally run on the modern versions of QEMU with no special customizations being required. In addition, Android comes with a driver for optical drives so existing workflows in EaaSI that allow for installing new software via an optical drive work with these versions of Android, also with no customization required. However, the configuration of QEMU to support desktop Linux-based operating systems and the way it is integrated within the EaaSI user interface assumes a limited number of inputs and outputs. Mobile devices generally have quite different input and output methods relative to desktop computers. For example, mobile devices often accept touch input, GPS sensor input, gyroscopic sensor input, and many other mobile-oriented inputs, and will output mobile-oriented outputs like vibrations, device-based-sharing protocols, and other mobile-specific outputs. Within EaaSI there is significant work to do to integrate these mobile-specific inputs and outputs into the EaaSI interface and workflows. In addition, we need to add options for simulating inputs like GPS coordinates and health sensor data.

II. TECHNICAL CHALLENGES

A. Emulation and Virtualization of Mobile Devices

The Android OS itself is an open-source project [3], however most Android versions that are used on handheld devices are modified by the device manufacturer to support device-specific operations. These devices overwhelmingly depend on the ARM(64) architecture. While ARM emulators exist,

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

emulating ARM-based Android images becomes a tedious task, as both performance and user experience suffer. Virtualization becomes a necessity if suitable user experience should be provided. As the desktop and server-based hardware on which the emulators run is usually x86(-64)-based, hardware acceleration and virtualization cannot be provided for images with full ARM emulation. Additionally, ARM emulation is more complex to set up as, in contrast to the x86(-64) architecture with the IBM PC platform, there is no universal ARM-based platform yet but many different platforms[1], i.e., combinations of a CPU implementing a specific instruction set architecture (ISA) together with other standardized hardware. While the ISA specifies the supported CPU instruction and their encoding, the platform allows other computer hardware to be expected to behave in documented ways and specifies, e.g., how the boot process works. As the IBM PC is the almost only x86-based platform, any x86 emulator will actually support it and, thus, can boot and run Android-x86. However, there is not yet any real equivalent for the ARM architecture (this is currently being fixed in https://www.arm.com/architecture/system-architectures/systemready-certification-program).

Given these challenges emulating ARM-based devices, there are currently two non-proprietary options that we've evaluated regarding ensuring long term access to the applications that were used on them:

1. *Android-x86*

Android-x86 is an open-source project with the goal of porting existing Android versions to the x86(-64)/IBM PC platform. The main advantage of Android-x86 is that it can be used with any x86-capable emulator, such as QEMU and works "out-of-the-box". Android-x86 is a community-based effort and thus there is no guarantee that the project will be continued and maintained long-term. Currently the latest release features Android 9, which was released in 2018. The current upstream Android Release is Android 13, with Android 14 being released this year. This shows that there is quite a discrepancy between the latest official Android release and that of Android-x86, however, with a long-term preservation view, this can be disregarded. Android-x86 allows to install the "Native Bridge" feature which allows to install and execute Applications within Android-x86 that were originally compiled to run on ARM-architecture only.

Most users of Android do not run the x86(-64) version.[2] Instead, they run a version of Android compiled for and compatible with ARM(64)-based hardware. ARM hardware has traditionally been more power-efficient than x86 hardware and so has been the default option for mobile devices that have limited battery capacity. Upon initial examination this might be expected to cause significant issues for preserving access to mobile applications as it would be reasonable to assume that most mobile applications were made for ARM-based devices. However Android and mobile applications each have some beneficial features that make this less of an issue that might be expected. From the beginning of mobile development many apps have been designed for either web-browser based execution or for use with a Java Virtual Machine (JVM)[3]. The web-browser based apps often work on any version of Android as they only rely on the in-built web browser. The Java-based apps will run on any version of Android as their architecture-independent bytecode is transparently compiled to machine code on the respective device itself.

2. *Google Android Emulator*

Google provides an official Android Emulator as part of Android Studio, the official Development Environment for Android. This emulator can also be used in standalone mode and consists of a QEMU with additional features and a somewhat complex emulator architecture [2]. Google provides Android system images, featuring different sizes, resolution, and most importantly Android versions. Images are provided in both x86 and ARM versions and range from Android 1.5 to Android 13. The Google Android Emulator offers a variety of input methods to simulate "real" input that a user would provide to a handheld device. It also contains interfaces to specify

[1] Historically, this has often been the case for CPU architectures, considering that the MOS Technology 6502 CPU and its variants were used in such diverse computers (platforms) as the Commodore C64, Apple II, TRS-80, BBC Micro, Super Nintendo, and many others.

[2] However, Google is recently providing an x86-based virtual Android environment on most of their Chromebooks.

[3] Implemented as Android Runtime (ART) on Android (starting with Android 5.0) and being one of core parts of the Android operating system.

locations, use phone services such as simulating incoming calls or messages etc.

From a user standpoint, the Google Emulator seems like the obvious choice to emulate Android especially regarding input. However, there are strong arguments against the usage of the Google Emulator for long-term preservation:

1. As mentioned above, though being built on QEMU, the Google Emulator cannot easily be integrated with the EaaS framework.

2. Maintainability cannot be guaranteed as Google often discards projects and not all of the emulator's functionality is publicly documented well

3. The Google Emulator mainly uses the Android Debug Bridge (ADB) and wraps commands in its interface. If the EaaSI platform is extended to support ADB anyway, the better solution is to use Android-x86 and the "normal" QEMU and build interfaces for ADB functionality as we see need.

B. User Experience

C. Application Installation Workflows

1. The Standard Application Installation Workflow

The usual "workflow" when installing an application is the following: The user opens the Google Play Store, selects the Application that they want to install, and clicks "Install". While this option currently works within emulated Android-x86, it requires both Internet connectivity (which can be provided by Emulation as a Service (EaaS)) and a Google Account. Regarding long-term preservation however, we cannot assume a functional App Store and thus cannot rely on the Google Play Store. Additionally, for security reasons, Google enforces a policy where apps need to target recent Android versions or, otherwise, will not be available through the Play Store.[4] Android Packages (or "APK"s), however, can easily be installed manually, either from within the OS, or externally via a remotely executed command, using ADB.

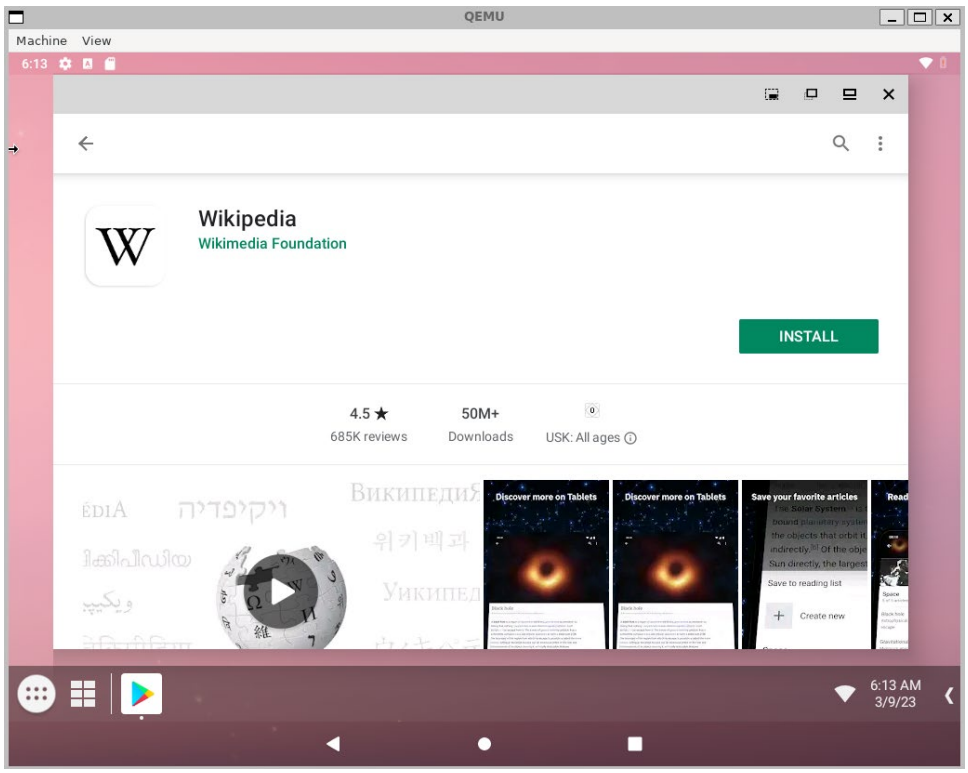


Figure 1: Installing the Wikipedia App within an emulated Android 9

4 https://android-developers.googleblog.com/2017/12/improving-app-security-and-performance.html

2. Side Loading Applications

By default, and in contrast to Apple's iOS mobile device Operating System (OS), the Android Operating Systems have allowed "side-loading" of applications once a user enabled the relevant system-setting. "Side-loading" is the process of installing an application from a file accessible to the device (e.g., downloaded from the internet, copied on removable media, or accessed from a network location). Apple does not allow this by default as they state that they consider it a security risk [4]. Side-loading in iOS is possible if the operating system and device are "jail broken". However, the process of "jail breaking" a device has many copyright and security concerns associated with it, this is one of a number of reasons why the EaaSI team has begun working with Android instead of iOS to address long term mobile application preservation and access.

The ability to side-load applications in Android OSes provides a simple and future-proofed method for preservation practitioners to use to ensure preserved mobile applications can be installed and accessed by future users. It is by taking advantage of this option that EaaSI users can use the existing software installation workflow to install applications in Android-based mobile devices emulated in EaaSI. EaaS currently supports installation through the following workflow: The user uploads that APK that they want to install. The backend wraps the APK in an ISO file that can be inserted into the emulated Android system, similar to how an SD card is inserted into a real smartphone. The user can then install the APK from the Android File Browser.

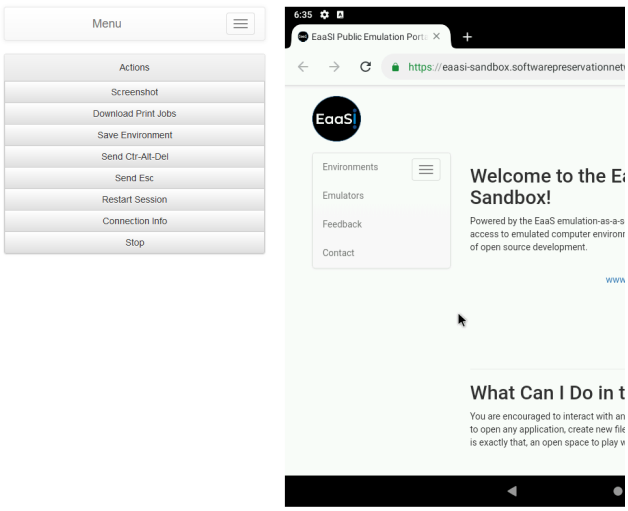


Figure 2: The Google Chrome Browser running within Android 9 in the EaaS UI

3. Providing a Custom Legacy App Store

Most users install applications in their mobile Operating Systems (OSes) by finding them in the OS's application store ("app store") and clicking the button to install the application. One option for managing workflows for installing software in preserved and emulated versions of mobile devices and their OSes would be to replicate this process, i.e., the EaaSI software could provide a server hosting the applications and an "app store" application on the emulated devices that would provide a similar experience to the app stores users are used to. This option is relatively complex and requires a network to be setup between the app-store host server and emulated device along with sufficient metadata to be populated into the app store database to enable meaningful searching and browsing within it. Such a configuration would also need to be maintained for as long as the need to install legacy applications was required. For these reasons, while the EaaSI team may explore this option in the future, we have instead decided to start with simpler workflows for enabling applications to be installed on emulated mobile devices.

4. Automatically Installing Applications

In addition to side-loading applications via the ISO-wrapping workflow described above, the EaaSI team have prepared Android-x86 images that automate this process using a startup script that checks if any APKs are present within the mounted ISO and installs them via the integrated 'package' command line tool. That means that no user input is required, and the installed app(s) can be used shortly after startup. This automated installation is possible because APKs usually don't require any external dependencies and can "just" be installed.

There are currently plans to expand the EaaS-Framework to allow external installation of APKs as well. As networks are already an established functionality in EaaS, a container with ADB could be connected to the Android Emulator. ADB could then be used to remotely install APKs over the network.

D. Interaction

Interacting with Android applications (or applications for handheld devices in general) differs from interaction with a computer. This poses new challenges for emulation as well. While many applications can run properly in an emulated environment, input and output in the form of multi-

touch is not supported in EaaSI yet. While the Google Emulator provides an interface to simulate Inputs, the current EaaSI-UI does not support any of these Android-related features. The above-mentioned network-based ADB integration will set a base for the integration of some of these features, where a UI input would be translated by the backend to an ADB command – which is similar to the way the emulator provided in Google's Android Software Development Kit (SDK) functions.

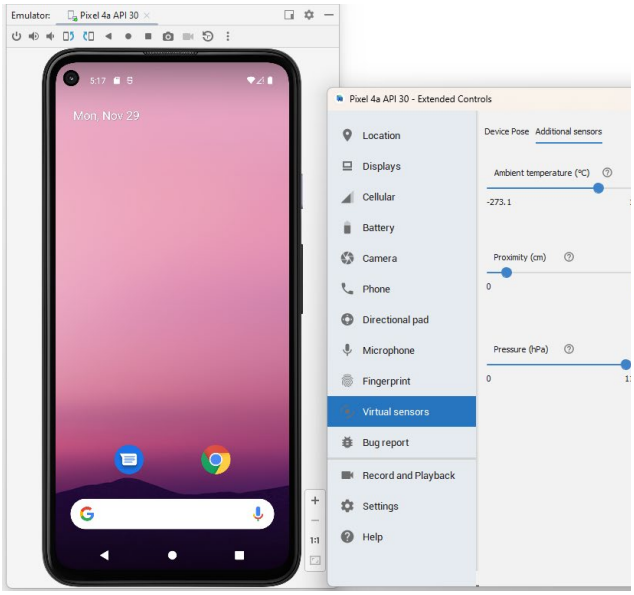


Figure 3: An example of the Google Android Emulator with extended Controls to simulate sensor input

E. *Virtualizing ARM devices*

While we have made great progress with virtualization of x86-compatible mobile devices, virtualization of ARM devices will require some additional work. Currently all servers that run EaaSI are x86(-64)-based and to virtualize ARM devices we would need to incorporate ARM(64)-based servers alongside the x86-based ones and develop EaaSI to support seamlessly utilizing and interacting with each type depending on the device being virtualized. ARM emulation on x86-based hardware will be essential in the future and is already possible. Given the delay between creation of born-digital archives and their acquisition by archival institutions (often many years), it may not ever be necessary to virtualize ARM-based devices in EaaSI as the existing emulation functionality may be performant enough on modern hardware to negate the need to virtualize the ARM devices.

F. *Apple device emulation and virtualization*

As discussed, we are not yet integrating any Apple device emulation or virtualization into EaaSI. There are some existing companies providing Apple device virtualization support, and some that claim to support Apple device emulation. However, in both cases their products are proprietary, which makes integrating them into the fully open-source EaaSI platform particularly challenging. In addition, there has also been significant litigation by Apple against these companies. The Software Preservation Network supported one company, Corellium (a company that supports security testing, training and research), in one such lawsuit by providing an amicus brief [5]. Fortunately, Corellium had some success with that case, however the risk of litigation is still high, which provides motivation for the EaaSI team to delay additional investment into integrating Apple device emulation and virtualization.

2. CONCLUSION

The EaaSI team have made significant progress in ensuring long-term access to mobile applications. A large proportion of historic Android-compatible applications can already be made accessible in EaaSI using the QEMU emulator and the Android-x86 operating system. Automated installation of these applications is also possible in EaaSI for many versions of Android. In the future, the process for installing and configuring these applications will be further simplified and streamlined using the Android Device Bridge. Additionally, configuration of and interaction with the more complex inputs and outputs that mobile devices and their emulators support is going to be integrated into the EaaSI User Interface enabling simple replication of the experience of interacting with legacy mobile devices and their applications.

Challenges remaining to be resolved include emulation and virtualization of Apple's mobile devices, however progress coming from the security testing industry seems to show promise for providing solutions that we may be able to integrate into EaaSI in the future.

3. REFERENCES

[1] Android-x86, Online. https://www.android-x86.org/

[2] https://source.android.com/docs/setup/create/avd, https://android.googlesource.com/platform/external/qemu/+/refs/heads/emu-master-dev
[3] Android Open Source Project, Online. https://source.android.com/
[4] Clover, J. 26/5/2022, MacRumours.com “Apple Says Revised U.S. Sideloading Bill Would 'Undermine the Privacy and Security Protections' iPhone Users Rely On”. https://www.macrumors.com/2022/05/26/apple-statement-revised-sideloading-bill/. Accessed 3/9/2023
[5] Butler, Brandon. (2022, February 17). SPN amicus brief defends fair use in Apple v. Corellium case. Software Preservation Network. https://www.softwarepreservationnetwork.org/spn-amicus-brief-defends-fair-use-in-apple-corellium-case/

SAVING STAN:

Preserving the Digital Artwork of Joseph Stanislaus Ostoja-Kotkowski

Taryn Ellis
State Library of South Australia,
Australia
taryn.ellis@sa.gov.au

Abstract — It is not unusual to find at-risk obsolete carriers in archival collections, but these 3½ inch ADFS-formatted floppy disks hold original digital artworks from the late career of a pioneering Australian multimedia artist. The graphics files, created on an Acorn Archimedes in the late 80s and early 90s, had not been seen for more than 25 years, and the difficult process of preserving and providing access to these artworks, and their associated software, highlights the fragility of the material from this era. The case study presented here discusses how the State Library of South Australia combined open-source and community software to automate the extraction and migration of obsolete content from these disks while capturing filesystem and other metadata—and discovered that emulation is not always the simplest solution.

Keywords — automation, migration, RISC OS, born digital, digital artwork

Conference Topics — From Theory to Practice

I. INTRODUCTION

The State Library of South Australia (SLSA) holds an extensive collection of material relating to the life and practice of Polish-Australian artist Joseph Stanisław Ostoja-Kotkowski, 1922-1994. The importance of this archive was recognized in 2008 when it was inscribed in the Australian Register of the UNESCO Memory of the World [1]. Among the 29 linear meters of material that makes up this collection are 940 3½ inch floppy disks containing original digital artworks, and related software, created in the late 1980s and early 1990s on an Acorn Archimedes computer.

The purpose of the project described in this paper was to preserve the data stored on these superseded, at-risk magnetic media, and to make that material visible via contemporary operating systems. Alongside this, the metadata captured during these processes facilitated the creation of thousands of catalog records, allowing the public to discover this collection and to access it online. Through our participation in the Archiving Australian Media Arts (AAMA) research project [2], we hope to join other cultural institutions in raising public awareness of, and increasing access to, Australia's born-digital cultural heritage.

Processing the 940 floppy disks was undertaken in three stages: (1) disks that had previously been assigned archival numbers, (2) unassigned disks, and (3) disks that were initially unable to be imaged. These stages roughly correlated with the disks' contents, with the first group containing mostly image files in an obsolete format; the second, a mix of software and images; and the third, disks that had been formatted differently, or that were of a different type. This last set also contains mostly image files, but in more common or contemporary formats.

II. STAGE ONE

A. A Slow Start

The 3½ inch floppy disks at the center of this project arrived at the library upon the artist's death in the mid 1990s. Their presence was noted in the catalog, and some were assigned accession numbers, but relatively little is now known about how they were processed.

An assessment undertaken in the late 2010s identified these as primarily double-density ADFS-formatted disks. A floppy disk drive and disk-imaging program [3] was purchased along with a commercial

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

Acorn Computer emulator [4]. The project was then revived in the context of SLSA joining AAMA in 2019-2020. The first 20 disks in the series were imaged and examined. From this sample it was confirmed that the subset of disks that had previously been assigned accession numbers predominantly contained digital artworks in the form of sprite files—a graphics file with extension 'ff9' that is native to RISC OS, and not recognized by contemporary software. Like the sprite files used in video games, these can contain one or more images, known as sprites [5]. Before work on the project could begin, Covid-19 arrived and everything stopped… and then started again.

The original project plan involved imaging each disk, mounting the image in the emulator and migrating each file individually to the PNG format via the RISC OS image-editing program, *Paint* [6], with renaming steps at intervals along the way. This was labor-intensive and risked accidental modification of the files in the editing software. We also observed issues with some of the migrated PNG files, which appeared strangely elongated when viewed via Windows. Further investigation uncovered that this ungainly aspect ratio results when contemporary software fails to correctly handle the rectangular pixels that comprise a subset of the sprite files. It is interesting to note that when this initial workflow was developed the organization was using Windows 7, which was able to correctly display rectangular pixels, a capability that was dropped in the development of Windows 10.

We needed to rethink our process.

B. Towards Automation

Online research revealed that academics, computer scientists, and RISC OS and Archimedes enthusiasts have been working on various projects both to extract data from RISC OS disk images and to migrate the sprite format to contemporary equivalents—to varying degrees of success. Not all of these shared tools were open-source or written in a common language like Python, and this made testing more difficult to carry out in a security-compliant manner. As well, timelines on the project were tight as resources had been allocated based on the original project plan, which we were now frantically re-working.

From the tools tested, we selected two community-generated programs—one to extract files from the disk images [7], and another to migrate the extracted sprite files to PNGs [8]. These programs were then incorporated into a custom Python script that controlled an automated workflow. In the event of errors—for example if the script was unable to extract files from the disk image, or if the correct number of PNG files were not created—the disk image (and any associated files) was automatically directed to a relevant 'problem file' directory for hands-on investigation. A second script was developed to manage metadata collection, taking text files generated in the first process and adding the contents to a spreadsheet. A project spreadsheet also served to capture information we could not extract programmatically, such as the disk brand or whether the disk label contained artwork. As well, we recorded if a disk failed the imaging process, and that disk was set aside.

C. Problem Solving

All was progressing well until we reached disk 161. We noticed a distinct color shift in the migrated PNG files: they were made up of pastel shades rather than the bold, saturated colors of the preceding files.

To save memory and disk space, some image files use an index or palette to centrally store color information, with each pixel pointing to this index rather than encoding detailed color information in place.

When we examined the 'pastel' sprite files using *Paint* in the emulator, we were surprised to find that the palettes for all these files were identical, but that they bore no resemblance to the colors as they were displayed on screen. When opened in the emulator, the 'pastel' files appeared to share the same color range as the earlier files.

With stabilizing this collection as our priority, we decided to continue imaging the disks while we had resources for this task, but to halt the file extraction and migration workflow while we reassessed.

This was also an opportunity to conduct a broader survey of the collection, and we used the emulator to examine disk images taken from across the 800 accessioned disks. In this way we encountered another conundrum—some of the

artworks, while still sprite files sharing the 'ff9' extension, were not recognized by our emulator. Attempts to render them in other emulators or via community-created tools produced mixed results: the programs that were able to render or migrate the sprite files produced very oddly colored and extremely elongated images.

We now had two, overlapping problems: files with color palettes our migration workflow was unable to interpret correctly; and file variations we were unable to view, let alone migrate.

In RISC OS the pixel resolution, pixel shape and color range of the display (desktop) is known as the screen mode and can be changed by the user. Sprite files also have a mode, and can only be displayed in that mode, or one very similar. Additionally, it is possible to create custom modes for files (and displays) that are not automatically supported by the operating system [9].

Examining the pastel files supported by the emulator revealed that they were always in mode 21, and it seemed likely that the files we could not view were in a non-standard mode or modes. We needed a way to programmatically determine a file's mode so we could identify files that were suitable for our current migration process and devise a new strategy for the others.

We examined sample files in a hex editor and compared the structure to online sprite format descriptions—a process which was complicated by the fact that there are variances between the different generations of sprite files and between our sprite files and those described. Using the byte offset of the sprite filename as an anchor point, we located the offset of the mode number. With this knowledge, we added code to our automated workflow to detect the mode and handle the file based on that, as well as to record the mode as part of our metadata collection.

By programmatically assessing each file's mode, we revealed that our collection holds sprite files in eight different modes. Four of these are proprietary modes, and all the files in six modes have either no color palette or an unexpected one. To manage the files in custom modes, we adapted a third program [10] so that it could recognize the distinct pixel resolutions and shapes represented by these modes and automatically output correctly formed PNGs. As well, experiments revealed that by programmatically assigning a standard 256 color palette to our mode 21 'pastel' files, we were able to output PNG files that were identical to those from the emulator, and so we applied this to the custom mode sprite files as well.

To judge the success of this tactic, we compared the PNGs migrated from the custom mode sprites to the thousands of as-yet undigitized transparencies that Ostoja-Kotkowski had taken of the artworks displayed on his computer screen. While mindful of the limitations of using these 'screenshots' to confirm color-accuracy, they nonetheless demonstrated that the PNGs we were producing to facilitate discoverability and access were an acceptable representation of the artworks. We added the third program to our automated workflow and resumed processing the disk images.

III. Stage Two

The second phase of the project was dedicated to preserving the 170 'unknown' disks, rumored to contain software. Initially the same process was followed: the disks were imaged, then the images processed via an automated workflow, with files and metadata extracted. For this subset we used a modified script with the sprite file migration streams removed.

We observed that scattered through this batch were more disks that solely contained artwork, and we wrote a script to find and move those disk images for reprocessing via our original workflow. We then sorted the remaining ff9 files by size to determine if any artworks had been overlooked and processed those that matched our specifications. To our surprise, we were able to identify a few images from Stage One, and began to question whether these files were, in fact, created by Ostoja-Kotkowski.

It was clear that we needed to look a little more closely at this material: we had the file lists that named all the programs and their component parts (including those curious image files), but we did not understand how the material was used or how grouped material was related.

By mounting the disks images in the emulator we made some important discoveries: (1) some of the programs were 26-bit rather than 32-bit applications, which made accessing them difficult [11]; (2) a 'duplicate' application, present on several disks, was actually a series of animated electronic letters

addressed to Ostoja-Kotkowski; (3) it was confirmed that some of the sprite files we had understood to be original artworks were actually demo files included with software.

The animated letters were especially intriguing. Written by a software engineer who was tailoring applications for Ostoja-Kotkowski, they provide insight into the artist's work practices and equipment. As well, we noticed that many of the programs, even those 26-bit applications we couldn't currently open, contained 'Help' or 'Read Me' documents that listed the creator's name, company or contact details, and information about the program. We identified the RISC OS text files (extension 'fff') extracted earlier, and programmatically migrated these to PDF. Among other advantages, by decoupling this content from the emulator, we could increase the discoverability of the collection and provide broader access to this research material.

IIII. Stage Three

With the majority of the disks now imaged, we were able to turn our attention to those that had failed the imaging step or were flagged as problems during the processing workflow. Looking at the latter group, it was apparent that the disk-image file sizes were varied, while the successful disk-images were all 800 KB. Testing revealed that our disk imaging software reversed the checkboxes for error handling so that when 'Write bad sectors as 0xFF' was selected, the program instead 'Skip[ped] bad sectors'. After adjusting for this and re-imaging the affected disks, we were delighted to discover none of the disks was so badly damaged that the content could not be accessed and extracted.

The disks that could not initially be imaged fell into two categories: (1) high-density disks (with their distinctive additional hole opposite the write-protection tab); and (2) disks that were formatted for DOS. By covering the additional hole with sticky-tape, we were able to image the high-density disks as double-density ADFS disks and process as before. Given the fragility introduced by the initial writing process we were lucky this worked [12]. The DOS formatted disks were imaged as '.img' files and their content accessed via Windows.

V. Wrapping Up

A project to preserve archival material such as this does not conclude when the final disk has been imaged or file migrated.

The physical disks have now been photographed to capture the annotations, doodles, and notes on their labels. An archival model has been developed to reflect the status and relationships of the files, and thousands of catalog records generated—utilizing the metadata we recorded along the way. An ingest plan has been designed to guide the movement of all this material to SLSA's cloud-based digital preservation system Preservica. At the time of writing, not all of these tasks have been completed, but progress is steady.

Although the role of emulation in the project was more limited than initially anticipated, it will become all-important in providing access to some of the software and peripherals uncovered. We will need to develop strategies: for managing the technical difficulties posed by application dependencies; for training new users in how to interact with a very different computing environment; for sharing this material beyond the single configured machine in the lab. These questions are much discussed in the Preservation community, and now that Stan's archive is safe, maybe we can take some time to look at them too.

Acknowledgement

The Archiving Australian Media Arts (AAMA) research project is funded by the Australian Research Council.

References

[1] UNESCO Australian Memory of the World Committee, "Documenting Visual Arts in Australia: Archives of Joseph Stanislaus Ostoja-Kotkowski," *The Australian Register UNESCO Memory of the World Program.* [Online]. Available: https://www.amw.org.au/sites/default/files/memory_of_the_world/documenting-visual-arts-australia/archives-joseph-stanislaus-ostoja-kotkowski.html. [Accessed: June 2023].

[2] Archiving Australian Media Arts, "Archiving Australian Media Arts: Towards a method and a national collection." [Online]. Available: https://aama.net.au/ [Accessed: June 2023].

[3] J. Watton, *Omniflop Wizard*. (Version 3.0). Sherlock Consulting Limited. [Online]. Available: http://www.shlock.co.uk/Utils/OmniFlop/OmniFlop.htm [Accessed: June 2023].

[4] G. Barnes and A. Timbrell. *Virtual RPC Adjust SA* (Version1.7.1.0). 3QD Developments Ltd. [CD-ROM]. Available: www.virtualacorn.co.uk/products/vrpcadsa.htm [Accessed: June 2023].

[5] R. Amos, "Pixel Graphics and Paint," in *Graphics on the ARM.* Swadlincote, UK: RISC World, 2007. [Online]. Available: http://www.riscos.com/support/users/grapharm/chap08.htm#L0005 [Accessed: June 2023].

[6] Acorn Computers Ltd, Cambridge, UK. *RISC OS 3.7 User Guide*. (1996). [Online]. Available: http://www.riscos.com/support/users/userguide3/book2ab/e_2.html#marker-9-47 [Accessed: June 2023].

[7] T. McLaughlin, *py3adf* (Version 0.1.2). [Online]. Available: https://github.com/jarpy/py3adf [Accessed: June 2023].

[8] S. Huber, *SpriteViewer/SpriteConverter* (Version 0.3.0). hubersn Software. [Online]. https://www.hubersn-software.com/; https://www.huber-net.de/risc_os/ Available: https://stardot.org.uk/forums/viewtopic.php?t=15801) [Accessed: June 2023].

[9] Acorn Computers Technical Publications Department, Cambridge, UK. *RISC OS 3 Programmer's Reference Manual*. (1992). [Online]. Available: http://www.riscos.com/support/developers/prm/modes.html [Accessed: June 2023].

[10] D. Boddie, *The Spritefile module: spritefile.py; spr2other.py* (2005). [Online]. Available: https://www.boddie.org.uk/david/Projects/Python/Spritefile/ [Accessed: June 2023].

[11] This relates to the development of the ARM chip, which started out with a 26-bit address space but 32-bit CPU processing. See the following for some interesting tidbits: https://encyclopedia.pub/entry/history/show/78043 ; https://lowendmac.com/2018/the-arm-story-riscy-business/ ; https://heyrick.eu/assembler/32bit.html [Accessed: June 2023].

[12] M. Brutman. "Working with Disks: An intro to floppy disks and floppy drives." BrutmanLabs.org. http://brutmanlabs.org/Diskettes/Diskette_handling.html [Accessed: June 2023].

PRIORITIZING STORAGE MEDIA FOR DIGITAL ARCHIVING AND PRESERVATION

Leo Konstantelos
University of Glasgow
United Kingdom
Leo.konstantelos@glasgow.ac.uk

Emma Yan
University of Glasgow
United Kingdom
Emma.Yan@glasgow.ac.uk

Abstract – This paper summarizes our efforts to-date at Archives & Special Collections, University of Glasgow, to develop a methodology and tool for prioritizing archival processing of digital collections stored in physical storage media. We present the sources and process we used to develop the methodology, and outline the functionality of a prototypical tool to generate prioritization scores.

Keywords – digital archiving, legacy storage media, prioritization, selection and appraisal

Conference Topics – From Theory to Practice, Immersive Information.

I. INTRODUCTION

This paper summarizes our efforts to-date at Archives & Special Collections, University of Glasgow[1], to develop a methodology and tool for prioritizing archival processing of digital collections stored in physical storage media.

The motivation and requirement for this endeavor arose from our ongoing work on digital archiving, archival forensics, and preservation of large at-risk born-digital collections [12-15]. Since 2019, the University of Glasgow Archives have been maintaining a register of digital assets deposited and maintained in the collections as physical storage media. As the asset register continued to grow, it became evident that the micro-appraisal approach we had thus far followed to select storage media assets for processing on a case-by-case basis, which was based on predominantly empirical criteria for selection and prioritization, were insufficient to deal with growing volumes. A recurring concern was the further degradation of storage media, the majority of which being already legacy and obsolete, in potentially aggravating conditions, to the point that they became inaccessible.

These problems are by no means new. The fragility of storage media, alongside the need to establish methodologies for addressing the preservation of the digital objects they contain, have been signaled repeatedly [e.g. 1, 4, 5, 6]. Extant studies from collecting institutions internationally, and community/web resources, have provided insight into the longevity, average lifespan and susceptibility to damage of different kinds of storage media [2, 6-11]. The 'Bit List' of Digitally Endangered Species [3] incorporates, extends, and contextualizes many of the concerns expressed in these studies, in a resource that is maintained and reviewed by the global digital preservation community.

The goal of the work presented here has been to encapsulate the knowledge/guidance deriving from these resources into a methodology that aligns with our workflows [12-13]; and use this methodology as the foundation for a simple tool to generate a priority score for processing physical storage media, that takes into account both community practice and other evidence-based criteria.

II. METHODOLOGY

In defining the criteria for prioritizing digital archiving and preservation of physical storage media, we drew from the methodological approaches suggested in [2] and [5]; and our work was further informed by the issues identified in [6]. To align the methodology with our workflows, and to keep the criteria as succinct and flexible as possible, we agreed on the following assumptions:

[1] www.gla.ac.uk/myglasgow/archivespecialcollections/

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

- The majority of digital assets contained in the storage media we hold in our collections are unpublished and have been transferred to our care without detailed file manifests or other descriptive documentation as to their exact contents.
- The digital assets contained in the storage media are as valuable, important, or otherwise intrinsic to the rest of the collection that they belong to, until they have been appraised.
- The digital assets contained in the storage media are unique, there is no other copy of the contents other than that on the storage media we hold, unless we have specific information to the contrary in our records.
- The initial focus will be on storage media that we can currently hold in our collections and can process via our archival forensics capability; and will exclude such legacy media as punch cards.
- Environmental conditions in which storage media were stored prior to being deposited to the University Archives, will be considered as 'aggravating' unless we hold information to the contrary in our records.

Three storage media-specific criteria were deemed as the most important in prioritizing digital archival processing and preservation:

- Average lifespan of the medium, as indicated in the examined literature [6-11].
- Year of production of the medium, as a measure of longevity and obsolescence [2].
- Environmental conditions in which the medium has been stored *after* being deposited to the University Archives. These draw from information recorded in [3] and [4].

We used the classification and – to the most part – the terminology adopted by the 'Bit List' of Digitally Endangered Species [3] as a fourth criterion, so as to inform the methodology with community practice; and avoid duplication of existing effort.

We assigned a score from 1-5 for each criterion, as shown in Table 1.

Table 1. Criteria and scores for storage media prioritization

Bit List' of Digitally Endangered Species	
Classification	**Score**
Lower risk	1
Vulnerable	2
Endangered	3
Critically Endangered	4
Practically extinct	5
Average lifespan	
Lifespan	**Score**
1-3 years	5
3-5 years	4
5-10 years	3
10-20 years	2
More than 20 years	1
Conditions	
Conditions	**Score**
Optimal conditions	1
Good conservation practice	2
Minimal conservation practice	3
Some aggravating conditions	4
Mostly aggravating conditions	5
Year of production	
Produced	**Score**
Within the last 5 years	1
More than 5 years ago	5

We collated information from the studied sources to generate a list of storage media types that are currently help by the University Archives and assigned prioritization criteria to each medium (Table 2). The list summarizes our current knowledge on average life span and contemporaneity of storage media; and reflects the status identified in the 'Bit list' as of the time of writing. The conditions of storage were purposefully left out of this summary list, as they are bound to differ per medium – for instance, a current portable HDD may have been stored in either optimal or aggravating conditions and this can only be gauged on a case-by-case basis.

Table 2. Collated types of storage media, with prioritization criteria assigned.

Medium	Produced	Bit list status	Average lifespan (years)
Current internal HDD	Within the last 5 years	Vulnerable	3-5 years
Current internal SSD	Within the last 5 years	Vulnerable	3-5 years
Non-current internal HDD	More than 5 years ago	Critically Endangered	3-5 years
Non-current internal SSD	More than 5 years ago	Critically Endangered	3-5 years
Current portable HDD	Within the last 5 years	Endangered	3-5 years
Current portable SSD	Within the last 5 years	Endangered	3-5 years
Current optical media (CD, DVD, BlueRay)	Within the last 5 years	Endangered	5-10 years

Medium	Produced	Bit list status	Average lifespan (years)
Current magnetic tape	Within the last 5 years	Endangered	10-20 years
Current Flash storage (USB stick, SD card)	Within the last 5 years	Vulnerable	3-5 years
Floppy disk	More than 5 years ago	Critically Endangered	1-3 years
Non-current magnetic tape	More than 5 years ago	Critically Endangered	10-20 years
Cassette tape	More than 5 years ago	Critically Endangered	10-20 years
Iomega zip disk	More than 5 years ago	Critically Endangered	10-20 years
Non-current optical media (CD, DVD, HDVD, Laser disc)	More than 5 years ago	Critically Endangered	5-10 years
Non-current portable SSD	More than 5 years ago	Critically Endangered	3-5 years
Non-current Flash storage	More than 5 years ago	Critically Endangered	3-5 years
Locally managed network storage	Within the last 5 years	Vulnerable	10-20 years
Cloud storage (third-party)	Within the last 5 years	Vulnerable	5-10 years
Current locally hosted web resources (websites, online databases)	Within the last 5 years	Vulnerable	5-10 years
Current externally hosted websites (websites, online databases)	Within the last 5 years	Endangered	5-10 years
Non-current locally hosted web resources (websites, online databases)	More than 5 years ago	Critically Endangered	3-5 years
Non-current externally hosted websites (websites, online databases)	More than 5 years ago	Critically Endangered	3-5 years

Lastly, we developed a five-point priority score that is meant to indicate the time period within which digital archiving and preservation action should be taken (Table 3).

Priority score	
Score	**Priority level**
1	Low priority - action within 3 years
2	Low priority - action within 1 year
3	Medium priority - action within 6 months
4	High priority - action within 3 months
5	Extreme priority - immediate action

III. Prioritization Tool

The prioritization tool is a simple proof-of-concept, which calculates a priority score (Table 3) for each type of storage medium, based on the scores identified for the individual criteria plus the conditions that a medium has been held in (Figure 1).

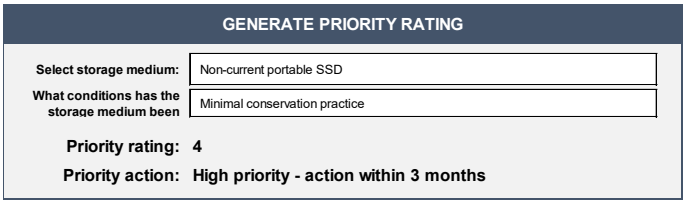


Figure 1. Screenshot of the storage media prioritization tool.

In its current version, the tool calculates a priority score using equal weights (25%) for each of the four prioritization criteria.

IV. Further Work

Being based to community-generated resources, guidance and practice, the prioritization methodology and tool are equally open to community feedback and discussion. Our aim with this piece of work is not to epitomize practice in this area, but rather invite dialogue and create a space for both further insights on handling computer storage media as archival records; and for reusing community-maintained resources, such as the 'Bit list'.

Recent discussions within our teams and with the wider digital preservation community on this topic, have highlighted issues with the score weighting. Specifically, it has been suggested that storage conditions should be given a higher weight, as it can adversely impact all other criteria.

The list of storage media that we have collated is neither complete nor comprehensive – and it is bound itself by obsolescence. Changes in community guidance, and the findings of future studies on the longevity and susceptibility of storage media, will require respective changes to the current scores. In this sense, it is an ongoing piece of work that provides the means to inform decision-making, rather than an end in itself.

REFERENCES

[1] Digital Preservation Coalition, “Digital Preservation Handbook: Legacy media”, 2nd Edition, 2015. https://www.dpconline.org/handbook/organisational-activities/legacy-media

[2] A. Brown, “Selecting Storage Media for Long-Term Preservation”, The National Archives, August 2008. https://cdn.nationalarchives.gov.uk/documents/selecting-storage-media.pdf

[3] Digital Preservation Coalition, “The 'Bit List' of Digitally Endangered Species”, 3rd Edition, November 2023. https://www.dpconline.org/digipres/champion-digital-preservation/bit-list

[4] A. Brown, “Care, Handling and Storage of Removable media”, The National Archives, August 2008. https://cdn.nationalarchives.gov.uk/documents/information-management/removable-media-care.pdf

[5] R. Erway, “You’ve Got to Walk Before You Can Run: First Steps for Managing Born-Digital Content Received on Physical Media”, OCLC Research, 2012. https://www.oclc.org/content/dam/research/publications/library/2012/2012-06.pdf

[6] D. Elford, N. Del Pozo, S. Mihajlovic, D. Pearson, G. Clifton, C. Webb, “Media Matters: developing processes for preserving digital objects on physical carriers at the National Library of Australia”, 74th IFLA General Conference and Council, August 10-14, 2008. https://archive.ifla.org/IV/ifla74/papers/084-Webb-en.pdf

[7] C.L. Erickson and B.M. Lunt, “Alternatives for Long-Term Storage of Digital Information”, Proceesings of the 12th International Conference on Digital Preservation (iPres 2015), November 2-6, 2015. https://digitalpreservation.lib.byu.edu/wp-content/uploads/2015/11/iPres_poster_2015_resubmit2-Erickson_Lunt.pdf

[8] R. Pinciroli, L. Yang, J. Alter and E. Smirni, "Lifespan and Failures of SSDs and HDDs: Similarities, Differences, and Prediction Models," in IEEE Transactions on Dependable and Secure Computing, vol. 20, no. 1, pp. 256-272, 1 Jan.-Feb. 2023, doi: 10.1109/TDSC.2021.3131571.

[9] B.M. Lunt, “How Long Is Long-Term Data Storage?”, IS&T Archiving Conference 2011 (Archiving 2011), May 16-19 2011. https://www.imaging.org/site/PDFS/Reporter/Articles/2011_26/REP26_3_4_ARCH2011_Lunt.pdf

[10] National Park Service, “Digital Storage Media”, *Conserve O Gram,* vol. 22, no. 5, October 2010. https://www.nps.gov/museum/publications/conserveogram/22-05.pdf

[11] Arcserve, “Data storage lifespans: How long will media really last?”, Arcserve Blog, July 2022. https://www.arcserve.com/blog/data-storage-lifespans-how-long-will-media-really-last#:~:text=Most%20hard%20disk%20drives%20(HDD,desktop%2C%20or%20an%20external%20HDD.

[12] Archives & Special Collections, “Digital archiving workflow (high-level)”, v0.6, University of Glasgow, November 2022. https://coptr.digipres.org/index.php/Workflow:Digital_archiving_workflow_(high-level)

[13] Archives & Special Collections, “Archival Forensics workflow (storage media deposit)”, v0.2, University of Glasgow, November 2022. https://coptr.digipres.org/index.php/Workflow:Archival_Forensics_workflow_(storage_media_deposit)

[14] L. Konstantelos, E. Yan and C. Paterson, “Integrating archival forensics with digital archiving workflows”, BitCurator Forum 2023, March 2023 (forthcoming).

[15] L. Konstantelos, E. Yan and C. Paterson, “Evaluating Digital Preservation Capability With Large At-Risk Collections. Lessons Learnt From Preserving The Nva Archive”, Proceedings of the 18th International Conference on Digital Preservation (iPres 2022), September 12-16, 2022, pp. 282-86.

TOWARDS PRESERVING WEB-BASED STUDENT PUBLICATIONS AT CONCORDIA UNIVERSITY

Sarah Lake
Concordia University
Canada
sarah.lake@concordia.ca

John Richan
Concordia University
Canada
john.richan@concordia.ca

Abstract – Student-run papers, journals, and magazines that were previously published in print form are now almost exclusively hosted on digital publishing platforms. How will this shift impact the longevity of student scholarship and institutional memory? This paper will present an ongoing project that aims to archive web-based student publications at Concordia University. We will discuss the rationale behind the project, our initial objectives and scope, and the challenges that we have encountered so far. We will conclude with a discussion of future opportunities for outreach and other envisioned pathways for collaboration.

Keywords – web archiving, student publications, university archives, academic libraries, Archive-It

Conference Topics – We're All in this Together

I. BACKGROUND

During the summer of 2018, the Concordia University Records Management and Archives (RMA) department started the process of web archiving within the greater context of its newly published Digital Preservation Program. With a departmental mandate directly tied to University records management activities, early efforts using Archive-It were focused on establishing and maintaining automated crawls within the Concordia domain. The objective of this work was two-fold: To preserve important information contained on Concordia websites and to respond to research needs originating from the community searching for information hosted on former University pages no longer accessible.

RMA web archiving has since evolved to respond to special interest topics, for example the Concordia COVID-19 Web Collection. With this evolution has also come the opportunity to collaborate with new partners and advocate for the importance of web archiving within the community.

One collecting gap that has been identified by RMA was archiving *online-only* student publications. These publications are characterized as being exclusively online journals or magazines managed by Concordia students showcasing undergraduate or graduate writing and work. Prior to the shift to online only, RMA collected extensively in this area and houses a large print collection of this type of material for research purposes. However, with the majority of this work now exclusively being published online, new technical and ethical questions have arisen for archivists and librarians.

In parallel to RMA's efforts, Concordia University Library began investigating the resource requirements of potential impact of developing its own web archiving program. A pilot project in 2021 uncovered collecting areas of interest, including websites related to Special Collections holdings and web-based scholarly output by Concordia researchers; and created a framework for an operational web archiving program at the library. With only one librarian with web archiving in their job description, ensuring the sustainability of the program is an ongoing challenge. In order to maximize the impact of the library's web collecting activities while balancing its limited resources, we continue to seek out opportunities to create impactful web collections that will benefit Concordia University and the broader research community. The project described in this paper is one example of such an opportunity.

II. PLANNING THE PROJECT

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

In 2021, we began to envision a collaboration between RMA and the library to collect web-based student-run publications and make them available using Archive-It. Concordia archivists and librarians recognized both the fragility and value of these websites, and believed that archiving them could have a meaningful impact. These publications contain unique scholarly output produced at the university and they hold important evidentiary value in documenting student life and culture. Archiving them would contribute to both RMA's mandate to preserve and provide access to the university's institutional memory and the library's mandate to preserve and provide access to Concordia research.

Except for two literary journals, none of the student publications we identified seem to exist in print form. This makes their contents even more at-risk, as web publications are inherently more fragile than their print counterparts and open access journals regularly disappear from the web [1]. Student-run publications are also particularly ephemeral due to their organizational infrastructure. They tend to be staffed by teams of student volunteers with high turnover, they operate with limited funding, and from what we can tell, most have no long-term stewardship or preservation plan in place. In the year since we started planning this project, we have already witnessed some of these websites disappear.

In 2022, we formed a joint working group which consisted of a small team of archivists, records managers and librarians involved in web archiving at the library and RMA. The working group developed the following plan for the project. We would begin by creating a seed list, i.e., a list of publications to capture and their URLs. We would then contact the editorial teams of these publications to notify them of the project, and begin to crawl the websites using Archive-It.

We expected running web crawls, reviewing results and troubleshooting issues to be the most time-consuming part of this project, especially since neither the library nor RMA has a single employee fully dedicated to web archiving. In order to expedite this process, we decided to enlist the help of a Library and Information Technician intern for spring 2023. Under the supervision of the project lead, the intern will run and review crawls and add descriptive metadata. Once we are satisfied with the results, our last step will be to make the collection publicly available on Concordia University's Archive-It page.

III. Challenges

Creating an initial seed list proved to be more challenging than we had expected. These publications tend to be siloed in their respective faculty websites which makes them difficult to find. Without a centralized directory, we had to rely on browsing the different department web pages to find their associated student publications, which were often buried many sub-pages deep. Compiling a full inventory of all these publications would be quite onerous, so instead we aimed to create an initial seed list that would represent a sampling, rather than an exhaustive list.

Determining selection criteria for publications was another challenge. Some journals were exclusively managed by students, while others were managed by a combination of students and faculty. Some featured only student work, while others featured work by faculty and authors external to the university. Some websites pushed the boundaries of what we considered web-based publications and we were faced with difficult decisions: should we include a student-run conference website, a scholarly podcast, or a fanzine produced by a special interest club?

We decided to start with a narrow scope to keep this first phase as simple as possible, with the expectation that it could eventually be broadened to include a wider range of publications. We limited our seed list to 16 online journals and magazines that were self-described as exclusively student-run and featuring exclusively student work. We kept a spreadsheet of the websites that we had considered but ultimately scoped out, with appraisal notes explaining our decisions.

Our next step was to contact the editorial teams of our selected publications to inform them of the project and seek their collaboration. This proved to be challenging as well. Many institutions consider an opt-out rather than opt-in approach to be the only workable solution to web archiving, in part due to the typically low response rate from site owners which makes it difficult to obtain explicit permission to archive content [2]. A 2017 survey by the NDSA, Web Archiving in the United States, found that 70% of surveyed institutions capturing content do not seek permission or attempt to notify the content owner that their website is being archived [3].[1]

At the outset of this project, neither the library nor RMA had a formal web archiving policy in place, and the project team was hesitant to make the captured content available without the explicit consent of the website owners. The library is currently in the final stages of drafting a public-facing web archiving policy that will include opt-out and take-down procedures, and we plan to make the captured publications publicly available once this policy is approved. For now, we decided to notify the editorial teams of our selected publications by email, explaining our intention to crawl their site and inviting them to contact us if they have questions or if they wish to opt out. We drafted an email template and kept track of which publications had been notified using a spreadsheet.

As expected, we received few responses and struggled to make contact with the editorial teams. In the cases where we did make contact, we were faced with the challenge of managing the editors' expectations about how their content would be archived. The editors of one journal initially misunderstood the project as a backup service, where captures of their website would be replaced and refreshed periodically. When we explained that the aim of the project was to preserve the content in perpetuity, they chose to opt out, as they wanted to retain control over the archived content and to be able to edit the captured versions of their publication. This exchange highlighted the importance of clearly communicating the objectives and scope of the project to the editorial teams and allowing them to make informed decisions about their participation.

Moving forward, we anticipate technical challenges in crawling some of the publishing platforms that these publications are hosted on. For instance, two of the publications on our seed list are hosted on Issuu, a digital publishing platform that is not easily captured and rendered by Archive-It. As of the writing of this paper, Archive-It's help guide states that while they are working on improving their ability to both capture and replay Issuu publications, "at present, successfully archived Issuu publications

[1] As of the writing of this paper, the results of the 2022 edition of this survey have yet to be published. If the trend from previous iterations of the survey maintains itself, this percentage will have increased since 2017.

will not fully replay" [4]. The wide range of hosting platforms, each with their own technical particularities, means that post-crawl quality assurance and troubleshooting could prove to be significantly time-consuming. We may even need to investigate providing alternative means of access to some captured content, such as that hosted on Issuu.

IV. Conclusion

With the transition from archiving traditional print-based student publications to publications exclusively hosted online, new challenges and opportunities for archivists and librarians at Concordia University have emerged. These unique publications with no print equivalent are often at risk of going offline or undergoing major transformation rapidly. Factors related to the high turnover of students involved in these projects between academic years (sometimes more frequently) are central to the haphazard management of these websites. These sites regularly contain unique scholarly output, as well as a glimpse into student life and culture. In turn, this type of material holds high research value.

Within the context of this collaboration between RMA and Concordia Library we have presented some of the non-technical and technical challenges associated with archiving student publication websites. On the other hand, these challenges have presented archivists and librarians new opportunities to engage with non-traditional archives users. The discussions and reflections brought on by this project have inspired us to envision and implement new outreach initiatives in the Concordia community. For instance, the library has started to offer regular web archiving workshops to empower students and faculty to preserve their own web content using free and open-source tools. In the last number of years RMA has promoted its web archiving efforts through Concordia-based articles, presentations, various blogs and social media channels.

Through these collaborative projects, continued advocacy and training it is hoped that web archiving at Concordia can foster a more engaged group of stakeholders in terms of preserving and making accessible this type of at-risk information.

1. REFERENCES

[1] M. Laakso, L. Matthias, and N. Jahn, "Open is not forever: A study of vanished open access journals," *The Journal of the Association for Information Science and Technology*, vol. 72, no. 9, pp. 1099-1112, 2021. Available: https://doi.org/10.1002/asi.24460.

[2] C. Davis, "Archiving the Web: A Case Study from the University of Victoria," *Code4Lib Journal,* 26, 2014. Available: https://journal.code4lib.org/articles/10015.

[3] M. Farrell, E. McCain, M. Praetzellis, G. Thomas, and Thomas, P. Walker, "Web Archiving in the United States - A 2017 Survey," National Digital Stewardship Alliance (NDSA), 2018. Available: https://doi.org/10.17605/OSF.IO/3QH6N.

[4] M. Praetzellis, "Archiving Issuu and Scribd," Archive-It Help Center. https://support.archive-it.org/hc/en-us/articles/208333043-Archiving-Issuu-and-Scribd#HowtoarchiveIssuupublications (accessed Feb. 28, 2023).

(How) It Works!

A manifesto .. towards establishing a functional software collection at the Vienna museum of science and technology

Nika Maltar
Technical Museum Vienna
Austria
nika.maltar@tmw.at

Almut Schilling
aBITpreservation
Austria
almut@abitpreservation.net
0000-0002-2091-1579

Abstract - This paper presents the decision making process involved in establishing a software collection at the Vienna Museum of Science and Technology. The museum's collecting activities have been limited to the collection of tangible heritage. The current collection strategy defines the functionality of a museum object solely as its own material manifestation. That is why the museum keeps its physical collection in a "powered-off" state to preserve its integrity and functionality for the future.

To integrate a functional software collection into this theoretical frame we are discussing applied terminology and have developed a manifesto to build on a solid theoretical foundation.

Keywords – software collection, strategy, manifest, open source, embedded community

Conference Topics – Digital Accessibility; From theory to practice

1. Introduction

The Vienna Museum of Science and Technology houses one of the largest and oldest collections of technical objects, inventions, designs and research projects from various fields who have contributed to the advance of science, art and daily life of Austrian people. The largest part of the museum collection consists of commercial objects that were mass produced. However, the museum also preserves individual objects, art, innovations and technical inventions that were never mass-produced but are of particular value to Austria's scientific and cultural heritage. The various objects are collected and divided into five collection groups, each with its own individual collection strategy and research focus. The recent collection strategy for tangible cultural heritage protects the collected objects in the museum and prohibits any functional use of them to preserve their integrity.

Despite the diversity and historicity of the museum, which took on the task of preserving the physical integrity of the objects – keeping them in a "conservatorial resting state", its role as a collector of modern technologies and intangible cultural heritage in Austria was unclear. Due to the threat of technical obsolescence and the associated loss of the logical counter-part of the objects that were still undiscovered within the collection, as well as the growing need to digitally expand the museum's collection, the museum established a new collection department for the intangible cultural heritage as part of the research institute in 2022: The software collection.[1]

In addition to the obvious question: how to adapt the museum's existing infrastructure and collection strategy to meaningfully collect, catalog, document, preserve and restore the logical, the contextual and physical layers as a whole; there was also a conceptual issue. How to expand the internal methods and infrastructure with novel tools, technology and platforms to apply preservation actions on all described levels of threat. Does it make sense to build a software collection without planning the functional preservation of the original hardware to interpret?

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

II. Scope of the Research

At the beginning of this research project, the museum successfully applied for funding from the Vienna Business Agency and set a time frame of two years (12/22-12/24) to develop a concept for collecting, archiving, documenting and the dissemination of software-based objects, focused on two different groups: 1. complex software objects, 2. industrial everyday technologies with embedded software. The first group lays a focus on experimental games, design and computer graphical methods from the 1980s, 1990s and 2000s. The second group focuses on contemporary everyday machines from the museum such as the ATM machine [2].

First objective was to understand how we can utilize the existing physical collection, to preserve and extract the original source code, the executable binary and other functional software and hardware dependencies that in order to document their authentic performance, such as: applications, compilers, software libraries, operating systems, device drivers, firmware, hardware embedded software, etc.

II.A Focus

Since this project is right at the beginning this paper tends to openly discusses terminology used and the results of the ongoing inventory analysis of the existent collection and the associated decision making processes within the museum transformation. First step was: finding out what the profile of the collection is and how to logically expand it with device, resp. object relevant software and cultural-historical digital artifacts to *"reflect upon the development of technology and science*"[1].

II.B Goals

At the end of this research the following goals should be accomplished:

> Enhance the existing collection strategy with modified definitions and unify vocabulary and terminology.

> Integrate the new objects logically and conceptually into the collection;

> Establish a common understanding within the museum what "function" means to consequently preserve it properly.

> Build and integrate a sustainable, functional, and long-term software archive.

> Build a dedicated workspace with an emulation framework for hardware embedded software and data extraction, migration and rendering.

> Make this acquired knowledge accessible to other researchers, institutions, collection, archives and museums following the open source, open data, and participatory collaboration policy of the museum.

II.C Research Questions

? What kind of objects are already part of the collection (technical and historical)?
? How many of them are unique?
? Can a specific focus be deduced? (office, art, game culture, ...) and shall this focus be followed?
? What are the consequences for the software collection?
? What general strategy and focus can be derived and defined?
? What gaps need to be filled based on the developed strategy?
? How to define and apply the terms *software, information, digital entities, complex digital objects and their interrelations*?
? How to define this new group of objects within the context of this specific collection?
? How to re-enact the historical context of the object with its digital (virtual) twin?
? How to identify valuable content?
? What virtualized existence should we preserve?
? How to identify cultural heritage institutions with similar collection profiles and compare their existent content?

2. Related Work

Through an environment analysis we defined the following types of cultural heritage institutions, to engage within the international trans-disciplinary collaborative network: Technical museums, Computer (game) museums, Archives / Libraries and Art Collections. This pool of diverse scientific

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

DOI: 10.1145/nnnnnnn.nnnnnnn

disciplines frames our field of interest. As we tend to understand a museum artifact as complex entity which needs to be re-interpreted again and again to be perceived. This is caused by the inherent dynamic existence of the digital object itself and its rendering environment. In the case of the Technical Museum Vienna this rendering environment changes depending on the target group, stakeholders and use case (e. g. exhibition in the museum space, virtual access through online collection, etc.).

3. INVENTORY ANALYSIS

To comprehensively understand the physical collection: 199.338 objects / data sets (counted by inventory number) from the collection management system have been extracted, structured and classified, to:

- Understand and describe the technical and conceptual profile of the already existent software (either stored on information carriers or device embedded).

- Discuss this profile within internal and external stakeholders to enable a gap analysis and the potential expansion of the collection.

- Derive and define object groups of expansion considering technical and conceptual aspects (e. g. Austrian developed software, application software, games, art, external drives, ...).

- Identify and Execute risk assessment.

The following conclusions were drawn from the initial collection analysis: Interestingly, no focus on software (or its hardware environment) produced in Austria could be identified. The reason for this could be that the collection departments place a stronger focus on collecting objects with a cultural and technical connection to Austria (e. g. a series of generic desktop computers used in an Austrian bank branch). Furthermore we discovered that the found software was mostly hardware related (e. g. drivers and applications for an office printer) or were embedded in the collected hardware (e. g. an ATM machine). Based on the analysis, the historical context of the object as well as the context of their use along with its physical integrity are the main interest of the collection strategy.

Around 27% percent of the identified relevant objects are saved on different information carriers and embedded in dedicated devices. The rest constitutes itself as diverse hardware devices (personal computers, workstations, game consoles , digital music instruments, external reading devices). Most of the identified software objects are common computer and video games. We expect the 58% of unknown software (saved on different carriers) as mainly empty (~ useless), since the focus was to study the physical characteristics, rather than its intangible content. (see TABLE I)

TABLE I
classification of expecting software types

Type of software	**%**
Software Objects (Game, Doc, ...)	32
Supporting Software (Application)	3
Operating System	~ 1
Device Driver	~ 3
Hardware embedded	~ 1
unknown	58

4. TERMINOLOGY: INFORMATION, SOFTWARE, COMPLEX OBJECTS, FUNCTION AND PERFORMANCE

After the inventory analysis was completed, a general classification of the object groups was created and relevant keywords were identified. This made it possible to develop a more detailed description of the role of each object type and describe them based on their level of existence: the physical, logical and conceptual.[3]

> Definition on the physical level of existence (binary): All physical objects that contain transferable information or the logical part can be separated from the physical carrier and migrated into a virtual format. These binary images will form the basis of the software-archiv [4]: a passive, non curatorial selected collection of software objects to built up a functional infrastructure. Which means that they can be used in combination with other virtual binaries and environments. Their content should be extracted to allow precise interpretation. Based on the inventory analysis, we consider the following terminology to describe these types of software-objects: device-embedded software, hardware image, extracted virtual image, base

image, imaged medium, imaged system, synthetic image.[5]

> **Definition on the logical level of existence:** single binary or textual file, collection of binary or textual files, complex object with just internal dependencies, complex object with external dependencies or source code.

> **Definition on the conceptual level of existence**: software objects and supporting software (games and applications) operating systems, device driver, source code.

> **Definition of software-based objects**: targets are by definition digital born and digital transferred data objects [6] (binary and textual) as singular file or complex organization of files (dependencies).

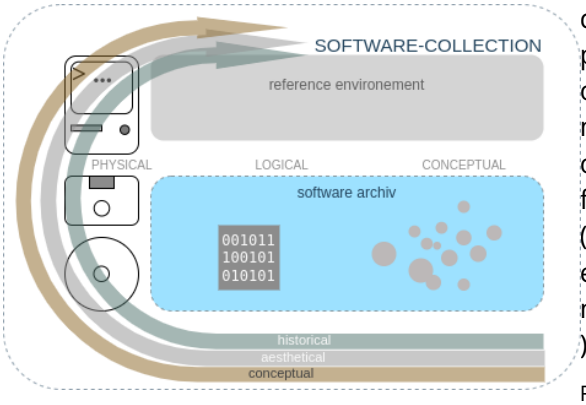


Figure 1 visualization of the structure of the software collection, consisting of the software archive, interpreted by the reference environment and representing historical, aesthetically and conceptual information.

> **Restructuring the existing tangible collection:** The primary objective is to first separate the physical and logical parts of the software-based objects in the Museum. We need to document the technical information, render them in the native environment, document their performance, and keep them in a stable, virtual form and archival formats. The second goal is to find a platform on which the separated virtual images can be merged into a working entity. The third goal is to permanently make the software archive accessible to researchers and the public.

We plan to build a software archive, that will contain the binary images and their extracted content. Re-interpretation and re-execution will be enabled by a reference environment, both in physical and virtual form (EaaS). While the virtual environment will provide access to the transferred software objects the physical environment will enable precise comparison with the originated hardware. Facing the museums politics the collections strategy has been modified to include the term “everyday objects” [14]: this excludes objects which are mass-produced, not older than 100 years, without any unique value from the restriction to “turn it on”. To substitute this collection donations will conclude the hardware environment and integrate the conceptual information (oral history) to the software collection.

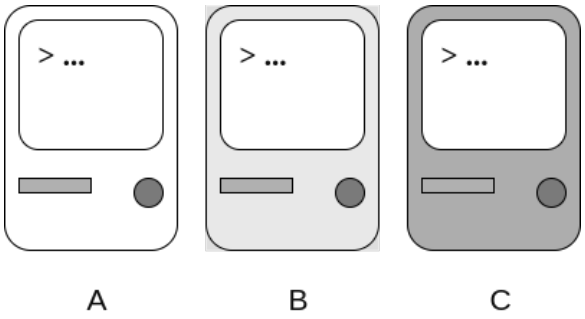


Figure 2 visualization of the different object/collection groups, resp. the defined environments and their associated use within the museum context. A: Objects (traditional museum objects). B: Functional Collection (non cultural heritage objects), C: Software collection (with software archive)

5. MANIFESTO

To create a profound understanding and clear communication of the software collection and its implications we drafted a manifesto resulting from the discussion described above. Considering the mission statement : "here technology becomes experience" we do state:

FUNCTION REFLECTS THE ABILITY OF A SIGNIFICANT PERFORMANCE: time is crucial for an authentic experience. The rhythm of processing and interaction explains the fundamental design and programming structure. Functional long-term archiving is focused on permanent access to digital objects and their interactive perception through a rendering environment. So we do understand the function of a rendering digital environment as a performative act revealing the truth of the object itself. This digital calculating space embodies more the dynamic structural relations than static hard coded numbers.[13]

NO INTEGRITY WITHOUT FUNCTION: This inherent functionality is key to the software object integrity.

The significance of performing is just possible through dynamically preserving the object, view path and interfaces to render (interpret) embedded instructions manifested in algorithms.

THE RENDERING ENVIRONMENT FORMS THE IDENTITY AND IS KEY TO DOCUMENTATION: As above stated software-based media objects need adequate environment to render it, not just for perception but also to document. Preservation implies documentation which requires just as much dynamism. The applied documentation methods are demanding permanent transformation otherwise they will get obsolete. So the primary operation of the archive is shifting from the content of one singular object to a logistical interlinking of object and layers.

DIGITAL OBJECTS DEGRADE INEVITABLE: Even though digital copies are identical, real life and endless (format) migrations disprove the myth of the digital … [8]. The Copying of data in the digital realm is questioning the meaning of the original but also blurs its boundaries. Lots of copies keeps stuff save but what is the difference between them? How to describe the different rendering environments and their significant properties? And even physical realities transform in time which might makes the concept of the static archival storage obsolete itself.

TECHNOLOGY IS CULTURE: The holistic claim of the historical museum object demands comprehensive investigation and documentation of the semantic level as well. The culture production from the 70s on has been especially driven by the development of these computing machines. The creation of video games is just one part. The computer as performing complex creates, embodies and transforms culture. The active protagonists of this scene are urged to research themselves to adopt documentation as cultural technique [7]. . . and preserve the process of the process of the process as permanent performance.

6. ACKNOWLEDGMENTS

This research has been funded by the WWTF Vienna Science and Technology Fund and the Vienna Business Agency.

7. REFERENCES

[1] https://www.technischesmuseum.at/museum/our_mission

[2] https://www.technischesmuseum.at/museum/online-collection#sammlung/ui/%7B%22search%22%3A%22diebold%20nixdorf%22%7D/objectdetail/600895

[3] National Library of Australia. 2003. UNESCO guidelines for the preservation of digital heritage. Information Society Division.

[4] Suchodoletz, D. 2008. Funktionale Langzeitarchivierung digitaler Objekte. Albert-Ludwigs-Universität Freiburg im Breisgau.

[5] Espenschied, D. 2017 MoMa Peer Forum

[6] Guttenbrunner, M. 2014. Establishing and Verifying Authentic Performances of Digital Objects: A Framework and Process for Evaluating Digital Preservation Actions. PhD thesis, Technische Universität Wien.

[7] Wark, M. 2004. A Hacker Manifesto. President and Fellows of Harvard College.

[8] Manovich, L. 2001. The Language of New Media. MIT Press.

[9] Briet, S. 1951. What Is Documenation? Editions Documentaires Industrielles et Techniques.

[10] Ensom, T. 2019. TECHNICAL NARRATIVES: ANALYSIS, DESCRIPTION AND REPRESENTATION IN THE CONSERVATION OF SOFTWARE-BASED ART. PHD Thesis. Kings College London.

[11] https://www.technischesmuseum.at/museum/forschungsinstitut/software_collection/die_panzerknacker

[12] https://www.technischesmuseum.at/museum/forschungsinstitut/software_collection/spielend_sammeln

[13] Ernst, W. 2013. Digital Memory and the Archive. University of Minnesota Press.

[14] Austrian Heritage Protection Law §2

REVISION-SAFE ARCHIVING AND LICENSE-CONTROLLED ACCESS USING DISTRIBUTED LEDGER TECHNOLOGY

Sven Schlarb
AIT Austrian Institute of Technology
Austria
sven.schlarb@ait.ac.at
0000-0003-3717-0014

Roman Karl
AIT Austrian Institute of Technology
Austria
roman.karl@ait.ac.at

Victor-Jan Vos
NIOD Instituut voor Oorlogs-, Holocaust- en Genocidestudies
Netherlands
v.vos@niod.knaw.nl

Carlijn Keijzer
NIOD Instituut voor Oorlogs-, Holocaust- en Genocidestudies
Netherlands c.keijzer@niod.knaw.nl

Begoña Sanchez Royo
Highbury Research and Development Ltd.
Ireland
begona.sanchez-royo@highbury.ie
0000-0002-2911-7881

Abstract – This paper describes an approach and a prototype system to make use of Distributed Ledger Technology or, more specifically, Blockchain, to build a trusted digital repository with a transparent and traceable change record for events related to the preservation or action of requesting or granting access to digital information objects. The approach focuses on a notary use case where the information stored in the blockchain serves as a proof of evidence regarding the existence and integrity of digital information objects.

Keywords – Blockchain, Distributed Ledger, Digital Repository, Electronic Archiving.

Conference Topics – Digital Accessibility, From Theory to Practice

1. INTRODUCTION

In this paper we present an approach together with a prototype implementation [1] which aims at increasing trust in electronic archiving and digital repositories by enabling a transparent and traceable change history of archival records using distributed ledger technology (DLT) or Blockchain systems.

Digital objects stored in a repository are subject to a life cycle, that is, there are events that modify the objects themselves or the metadata related to them. Trustworthy archiving means that these changes are recorded as events in a transparent manner and that it is clearly documented who initiated the changes and for what reason.

One of the well-known application domains of Blockchain is the so called "decentralized notary" [2, section 7]. The principle is that the piece of data, such as the fingerprint (or hash value) – demonstrating the existence and integrity of a document – is stored in the Blockchain together with a timestamp. It is decentralized because – depending on the security setup – any node can initiate transactions which are then processed through the distributed consensus and added to the Blockchain.

We present a use case related to negotiating access to digital objects where an applicant – a researcher from the repository's designated community or a general user – requests access. Further we describe that the basic principle can also be used to record digital preservation events.

The event metadata can be persisted together with the information resource. We use PREMIS, a widely used metadata scheme for recording preservation events, in combination with blockchain transactions to provide a transparent, auditable and tamper proof change history record.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

The prototype implementation [1] demonstrates the principle of using a blockchain notary for recording events related to accessing or preserving digital objects and is designed to make use of the use of the European Blockchain Infrastructure Service (EBSI) [3] which allows making use of API functions to build a transparent and tamper-proof provenance and change history record without the need to set up a dedicated blockchain service infrastructure. The European Blockchain Services Infrastructure (EBSI) is a cooperation of 29 countries and the European Commission. It is a private blockchain-based system with about 30 nodes which largely builds upon the Ethereum ecosystem with several smart contracts written in Solidity defining the core of the EBSI functionality. This private Ethereum network is not accessible directly from outside by users and external developers. Instead, there are several higher-level APIs as the only way to access the system from outside. Developers can use this API to write decentralized applications in a similar way as with interacting with custom smart contracts directly, but with the additional EBSI compatibility.

The paper's outline is structured as follows: Section 2 provides an overview of related initiatives and work. Section 3 elaborates on the fundamental methods for interacting with the blockchain. Section 4 delves into the implementation details of the access and preservation use cases. Finally, Section 5 concludes the paper by summarizing the key findings.

2. Related Initiatives and Work

A series of standards is relevant for trustworthy archiving: The Reference Model for an Open Archival Information System (OAIS) defines the requirements for an archive or repository to provide long-term preservation of digital information. Based on the reference model several initiatives produced recommendations for certification criteria related to trustworthy repositories, such as [4], [5], and [6]. Even though these publications address mainly organizational infrastructure aspects and do not address the technical means for building trust, they represent the general frame for building "accountable record-keeping systems" [5, p. 8].

The relevance of blockchain technology for archiving is reflected in the large number of publications related to this topic. Very close to the approach presented here is the model of a blockchain-based system to assist the process of long-term preservation of digitally signed records presented in [7] and [8] and the project ARCHANGEL [9]. The difference of our approach is that we present a generic use case applicable to any type of archive and propose a way to link preservation and access metadata with the blockchain registry.

3. Interacting with the Blockchain

To be able to interact with the blockchain, or more technically, to send a transaction that will be included in a block, i.e., writing data to the blockchain, an account on the blockchain system is required. This account can correspond either directly to a user or to an external system that administers multiple user accounts by itself. Additional to a digital account protected by public-key cryptography, a link to a person or an organization must be established. We will focus more on persons, "natural persons" in legal terms, as the relevant legal contracts are often concluded on this level.

The proof-of-concept implements a user management that does not include an official verification of an identity by the respective national agencies of EU states. But the architecture will be modular to allow such an extension without major code changes. A potential extension could integrate eID [10], a digital building block that was created as part of the Connecting Europe Facility (CEF) program, which takes care of cross-border verification of identities. Similarly, the European Self-Sovereign Identity Framework (ESSIF) built into EBSI can facilitate the generation of user accounts with verified identities based on official documents and make it usable with the rest of the EBSI functionality.

Giving access to data is also not a task that is typically solved only by a smart contract. First, because most blockchain systems have no sophisticated reading protection, which means that data is readable by default for parties with access to the system. But a blockchain is often not an ideal storage system for many kinds of data, because of its persistent nature and its reduced capacity and throughput. Therefore, dissemination information objects will clearly not be stored on the blockchain. Enforcing the rules of the blockchain will be done by a component outside the blockchain system. This component does not need a distributed architecture and can be located at the

archive. For multiple archives the component can simply be run as multiple instances, or an archive could use its own implementation if it wishes so. It is important to note that by doing so, we do not lose the advantages of the blockchain system. If an archive gave access to a DIP without having a legal contract established on the blockchain, this would be in its own control and responsibility. On the other hand, an archive that does not give access to an applicant even though a legal contract has been established would violate that agreement, which could be proven by the applicant.

Regarding the roles which are involved in the information access use case we define three entities when establishing a license agreement: Provider Entity (PREMIS metadata: agent), Applicant Entity (PREMIS metadata: agent), and Object Entity (PREMIS metadata: Information Package, Representation, or File).

The license agreement is concluded between the Provider Entity and the Applicant Entity, and it relates to the Object Entity.

For the preservation use case, events, such as the migration of representations of file objects, are documented as PREMIS with the corresponding agents documenting the preservation decision taken. The notary function of the blockchain registers a combined hash of event identifier and representation or file hash in this case.

4. Use Case Implementation

One of EBSI's APIs is particularly interesting for referencing data entries to prove their existence and validity at a certain point in time. The so called "Timestamp API" basically stores hashes and associates them to the timestamps at the point of time when the entry was created. The hashing algorithm can be chosen by the user out of a list of standardized algorithms.

The timestamp is added by the EBSI system when the entry is written to the blockchain. The original data can be stored off-chain. That might be possible also via the EBSI infrastructure or outside of it via a separate application.

An important aspect of the dissemination of digital objects is the definition of rights and the agreement concerning the usage. To define in which way and to what extent a dissemination object can be used, a legally binding contract between two parties, the provider, and a requester, must be put in place. In the context of systems that use a blockchain as basic data structure, there are pieces of code, called smart contracts, which are sometimes seen as an automated form of a legal contract. But this is only true in certain cases because smart contracts can only control very specific aspects of a legal contract. A smart contract cannot prevent the requester to use the dissemination object in any way that would violate the legal contract. But still, having a blockchain as data structure where it is not possible to modify existing entries helps us to digitally record an agreement between two parties and put the legally binding contract in place. The central part of the legal contract is the text that describes the legal aspects, which we call "license".

The process of recording the agreement on the blockchain consists of five steps:

1. Register dissemination representation for access. The dissemination object is identified and referenced to by a UUID. The dissemination object itself is never put on the blockchain, but only its identifier.
2. Create text license document. Most of the time, we will deal with a small set of standard licenses, but it is also possible to assemble a license out of a set of standard clauses or to set up an individual license. A license is hashed to prohibit future modifications and the hash is used as identifier for a license on the blockchain. Like the dissemination object, the license itself won't be stored on the blockchain.
3. Provider assigns license to dissemination representation. An entry on the blockchain is made to record the bundling of the dissemination object with a specific license. Everyone with access to the system will then be able to see the available offers. Since the blockchain contains only the identifiers, a customer will need not only the information from the blockchain to assess the offer. What is important, is that the data on the blockchain is sufficient for a customer to verify the validity of the offer.

4. Requester accepts license. The requester is the first to sign the offer via an entry on the blockchain.
5. Provider approves request. The signature of the provider via an entry on the blockchain finalizes the legally binding contract between one requester and provider for one dissemination object.

In these five steps we have three different fields identifying the different objects and the user (field: object identifier, type UUID, size: 16 bytes; field: license hash, type SHA3-256, 32 bytes; field: requester, type: Ethereum address, size: 20 bytes).

In steps 3 to 5, we need to store records which consist of a combination of the fields. If we want to use the Timestamp API, we cannot store multiple fields, but just a single hash value. We compute such a hash value by appending the input bytes in binary format and by then hashing this byte sequence with SHA3-256. As a result, we get another 32 bytes sequence representing a data entry.

This reduction still allows a party to proof the validity of a particular entry at a given time under the condition that the original data is not lost. By itself, the data stored on the blockchain itself will be relatively meaningless, so it is important to see it only as one part of the process.

To demonstrate the approach, we developed a decentralized application that is based directly on Ethereum, but because of the modular design the connection to the blockchain system can be replaced without the need of rewriting code in the other layers of the component. The blockchain-based system is set up as a private network with proof-of-authority as consensus mechanism. Go Ethereum is chosen as software for the execution client and a block is created every 15 seconds. But it is important to note that we do not rely on a particular setup and most of the setup could be changed without affecting the functionality of the decentralized application.

Adhering to the concepts and definitions of the EBSI, we created a timestamp smart contract that resembles the Timestamp service provided by the EBSI but is a simplified implementation. It defines a data structure, a map, with hash values as indexes and timestamps in combination with the address of the creator as associated information. It is implemented in Solidity, which is the most widely used language for smart contracts in Ethereum.

Interacting directly with smart contracts is in general a bit tedious. Most of the time, decentralized applications, sometimes also referred to as *dapps*, are created to interact with smart contracts. In our case the decentralized application is a web server that provides a REST API as an easy way to access the functionality of the smart contract. The web server is written in Haskell, uses GHC 8.10.x, and includes amongst others the libraries *servant* for the REST API and *web3-ethereum* for the connection to the smart contract. To ensure modularity, the application consists of 4 layers, that built only upon the layer directly below, but not on the others.

1. The uppermost layer in the architecture describes the REST API including all elements that concern the web server. This includes the rendering of the responses, for which JSON is used as format.
2. The second layer defines the business logic, which means the five steps in our process of recording an agreement and the functionality for retrieving information.
3. The third layer is responsible for the storage and is specific to the chosen storage backend. In the current implementation, the functions of the custom smart contract are called to store a hash or retrieve the information, timestamp and creator, of a hash. This layer has to be replaced if the EBSI Timestamp API is used instead of the custom smart contract.
4. The lowest layer is also part of the storage and just needed for the direct connection to the Ethereum components and wouldn't be needed with EBSI. It makes all the functionality of the smart contract accessible via the *web3-ethereum* library, as this is a form of polyglot programming and requires some mechanism to connect the different environments.

In the following we briefly outline the REST API functions of the prototype implementation.

- *registerObject*: takes *object_identifier* as parameter and returns the object's registration hash value.

- *createLicense*: takes *license_hash* as parameter and returns the license's registration hash value.
- *assignLicense*: takes *object_identifier* and *license_hash* as parameters and returns the license assignment hash value.
- *acceptLicense*: takes the requester's account requester, the *object_identifier* and *license_hash* as parameters and returns the acceptance hash value.
- *approveRequest*: takes the requester's account requester and the *object_identifier* as parameters and returns the approval hash value.

On the other hand, querying does not involve transactions on the blockchain, and the requests are relatively fast. For any of the registered hash values for object registration, license registration, license assignment, license acceptance, and approval one can get the timestamp of its registration and the associated account address.

- timestamp: takes the registration hash value as parameters and returns the timestamp value timestamp and the creator's account creator.

An Ethereum network is a peer-to-peer network consisting of several nodes. There can be one or more instances of the decentralised application that connect to one node, and one instance can serve one or more users. In the prototype implementation the key store is located at the Ethereum node, but it is possible to relocate it to the decentralised application, which would increase the flexibility and depending on the setup can also help increasing the security. There are some variations depending on the location of the key store and the number of the accounts that are stored in one key store. An interesting case is created by locating the key store at the decentralised application with only one account per key store. This would mean that every user has to run its own web server. An exchange of the custom smart contract with the EBSI services has only a minor impact, as it is also possible with the EBSI to manage the keys either at the node or outside of it.

5. Conclusion

The implementation of the concepts and approach presented in this article can be used with Distributed Ledger Technology or Blockchain services which offer a function for registering a hash value and providing a timestamp as return value. These minimum requirements will allow tracing the creation and integrity of information objects. To also provide evidence for the authenticity of information objects, i.e., who registered them for the first time or who originated specific preservation objects, the events need to be linked to the account. If the requirement is to know about real identities behind the accounts, further identification services, such as the European eID services, need to be integrated.

6. Acknowledgement

This project has received funding from the European Union's CEF programme with action No 2020-EU-IA-0185 under grant agreement No INEA/CEF/ICT/A2020/2397190.

7. References

[1] Kark, R., & Schlarb, S. (2023). Blockchain Notary Proof-of-Concept (Version 1.0.0) [Computer software]. https://doi.org/10.5281/zenodo.8100250

[2] D. Di Francesco Maesa and P. Mori. Blockchain 3.0 applications survey. Journal of Parallel and Distributed Computing, 138:99–114, 2020.

[3] W. J. Buchanan et al., "The Future of Integrated Digital Governance in the EU: EBSI and GLASS." 2023.

[4] D. R. C. T. Force. Trustworthy repositories audit certification: Criteria and checklist. Research libraries group (rlg), Technical report, RLG-NARA, 2007.

[5] R.-O. W. G. on Digital Archive Attributes. Trusted digital repositories: Attributes and responsibilities. Technical report, OCLC, 2002.

[6] W. G. on Digital Archive Attributes. An audit checklist for the certification of trusted digital repositories. RLG-OCLC, 2005.

[7] Bralić, V., Stančić, H. and Stengård, M. (2020), "A blockchain approach to digital archiving: digital signature certification chain preservation", Records Management Journal, Vol. 30 No. 3, pp. 345-362. https://doi.org/10.1108/RMJ-08-2019-0043.

[8] Stančić H, Bralić V. Digital Archives Relying on Blockchain: Overcoming the Limitations of Data Immutability. Computers. 2021; 10(8):91. https://doi.org/10.3390/computers10080091.

[9] J. P. Collomosse et al., "ARCHANGEL: Trusted Archives of Digital Public Documents," CoRR, vol. abs/1804.08342, 2018.

[10] H. Strack, O. Otto, S. Klinner, and A. Schmidt, "eIDAS eID & eSignature based Service Accounts at University environments for cross boarder/domain access." 2019.

NDSA Levels of Digital Preservation: A Review in Terms of Trustworthiness of Digital Records

Özhan Sağlık
Bursa Uludag University /
University of British Columbia
Türkiye / Canada
ozhansaglik@uludag.edu.tr
0000-0002-1436-7431

Abstract – Records created in organizations that have archival value should be preserved for a long time, and to achieve this, digital preservation techniques are used. These techniques also contribute to the preservation of the trustworthiness of the records. In order to assess the situation of organizations in the implementation of their digital preservation activities, there is a need for an analysis tool. Many models have been prepared to meet this need. One is the Levels of Preservation (LoP) developed by the National Digital Stewardship Alliance (NDSA). The LoP provides guidance to organizations in their digital preservation activities. Therefore, it is thought that the LoP can be associated with trustworthiness which aims at long-term preservation of the records. This study examines the levels of digital preservation specified in the LoP in terms of the trustworthiness of digital records. As a result of this research, the goal is to provide the basis for a methodology for organizations wishing to assess their level of digital preservation and to align their digital preservation capabilities with trustworthiness. This study used document analysis as a qualitative research design. Both field observations and research show that organizations are not sufficiently aware of the level of digital preservation and trustworthiness. Then, the question of the study is "how the levels that are specified in the LoP can be associated with the trustworthiness". As a result of the study, it has been observed that the levels of digital preservation specified in the LoP can be used in the analysis of the trustworthiness of the records. It is expected that this study will raise awareness in the organizations to do a better job of preserving the records that have archival value.

Keywords – Digital records, digital preservation, trustworthiness

Conference Topics – Sustainability: Real and Imagined; Immersive Information

I. Introduction

Records created in the ordinary course of business functions that have archival value are preserved for the long-term. It is known that digital preservation techniques are used to successfully meet this requirement. Digital preservation is defined as the series of managed activities necessary to ensure continued access to digital materials for as long as necessary [1].

These digital preservation activities cause organizations to analyze their current situation. Therefore, organizations may need an analysis tool. If so, methods such as developing a maturity model, obtaining certification, and conducting an internal assessment can be used. These methods are also used in preservation of the trustworthiness of the records. Here, trustworthiness means possessing the characteristics that the records are supposed to have according to recordkeeping principles and law. As a matter of fact, various approaches have been developed in this regard, both in the academic research and in scientific field studies. Electronic Resource Preservation and Access Network (ERPANET) [2], Cultural, Artistic and Scientific Knowledge for Preservation, Access and Retrieval (CASPAR) [3], Preservation and Long-Term Access Through Networked Services (PLANETS) [4], Alliance Permanent Access to the Records of Science in Europe Network (APARSEN) [5], CoreTrustSeal [6], Go FAIR [7] and International Research on Permanent

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

Authentic Records in Electronic Systems (INTERPARES) [8] can be given as an example.

In evaluating academic research, it has been found that Basma Makhlouf Shabou [9], Devan Ray Donaldson [10, 11, 12], Mpho Ngoebe and Jonathan Mukwevho [13] and Özhan Sağlık [14] have conducted studies on trustworthiness. Shabou criticized trustworthiness in Switzerland, Ngoebe and Mukwevho in South Africa, and Donaldson in the US. Sağlık, on the other hand, examined the evidential value of electronically signed records created in Turkish ministries in terms of archival trustworthiness in his doctoral thesis.

In the corpus of the International Conference on Digital Preservation (IPRES), there are many studies that assess the existing digital preservation capabilities of organizations. Although the LoP is also examined in some studies [15, 16, 17, 18], it cannot be observed that digital preservation capabilities are associated with the trustworthiness of the records. In other remarkable studies, the authors' observations at various institutions were presented [19, 20, 21, 22, 23]. However, there is a need for guidelines issued by organizations such as associations to measure the digital preservation capacity of institutions with different materials. Because these guides are designed with the needs of institutions that have many different types of materials. NDSA LoP, DigCurV Curriculum Framework and Digital Preservation Capability Maturity Model (DPCMM) and Rapid Assessment Model (DPC) developed by DPC can be given as an example [1, 24, 25, 26]. Among these studies, the LoP prepared by the NDSA stands out as a tool for organizations wishing to establish a digital preservation program.

The LoP, which can be used as a tool for organizations wishing to assess their digital preservation capacity, has five different functions and four progressive levels. These functions are storage, integrity, control, metadata, and content. The services provided by the organizations in these functions represent levels 1 through to 4 [24]. These services can be associated with trustworthiness.

In this study, the functions in the LoP are examined in terms of the trustworthiness of digital records. It is aimed to establish a methodology for organizations seeking to assess their digital preservation capability and to overlap the functions in the LoP with trustworthiness. As a result of this, it is thought that an awareness can be created in organizations to better preserve the records that have archival value. The study adopted a qualitative research design and used document analysis; the studies on this topic have been critiqued.

Both observations and studies show that organizations are not sufficiently aware of the digital preservation capabilities and trustworthiness [14, 27, 28]. In these circumstances, the question of the study is "how the levels that are specified in the LoP associated with the trustworthiness?" As a result, it is expected that an awareness will be created in organizations.

II. NDSA LEVELS OF PRESERVATION

The Levels of Digital Preservation are a tiered set of guidelines and practices for preserving the digital content. Levels can be used both education and advocacy and planning and assessment. But Levels do not reflect a holistic program that includes policies and procedures. They focus primarily on the technological aspects of a digital preservation program. There are four progressive levels in five different functional areas that can also be used to assess an organization’s digital preservation capability. Functional areas are storage, integrity, control, metadata, and content. These functions are evaluated in four progressive levels (Know, protect, monitor, and sustain) [24].

Knowing, the first level of the storage, includes criteria such as keeping content in a stable storage and having at least two copies in separate locations. An example of a level of protection criterion is keeping at least three copies, with at least one copy in a separate geographic location. Tracking the obsolescence of storage is one of the of the monitor level requirements. Performing tracked obsolescence is one of the criteria for the sustain level.

Generating integrity information and then verifying can be given as examples of the criteria questioned at the first level of integrity. One of the second-level criteria is to back up the integrity information and store the copy of it in a separate location from the content. Verifying this information at regular intervals is one of the third-level criteria. An example of a last-level criterion is to replace or repair corrupted content when necessary.

One of the exemplary criteria of the control function at the knowing level is to determine which authorization is to be exercised by whom and how. It is recommended that these authorizations be documented at the protection level. At the monitor level, the maintenance of log records can be cited as an example. Periodic review of access logs is one of the criteria at the final level.

At the first level of the metadata function, one of the first criteria is to create an inventory of the content with their current storage locations. Storing metadata is one of the criteria in the second level. At the third level, it is questioned whether a decision has been made about which metadata standards to be applied. Applying the adopted standards is one of the criteria of the last level.

The latest function is content. At the knowing level, it is sought to document the essential characteristics of file formats and content by including how and when they were identified. One of the criteria that can be given as an example at the protection level is to verify the essential characteristics of file formats and content. It is aimed to monitor the obsolescence and changes in the technology on which content is dependent at the monitor level. At the sustain level, it is asked whether activities such as migration and emulation have been performed.

The guidelines in the LoP can be considered as a milestone for digital preservation. Therefore, it is possible to examine these guidelines in the context of trustworthiness.

III. TRUSTWORTHINESS OF DIGITAL RECORDS

Trustworthiness is known as the preservation of attributes such as the medium, the content, the author, and the context of the records. The law, diplomatic and history disciplines that work directly with records have also developed various approaches regarding to preserving these attributes and maintaining trustworthiness. It is noteworthy that trustworthiness is defined differently in each of these disciplines. For example, for legal trustworthiness it has checked whether a record has the characteristics specified in the legislation; it has also checked whether the authorization mechanism is applied, and whether procedures are established in the records management processes [14, 29, 30, 31]. Diplomatic trustworthiness evaluates whether the form elements describing the records' characteristics are found appropriately. The procedures are analyzed by criticizing the features such as carrier, content, form elements, actions and persons in the record, archival bond, metadata, and context. It also examines digital signatures, seals, features of hardware and the software used, logs, audit trails and database transactions [14, 29, 31]. Another approach is historical trustworthiness. Here, it is checked whether the information contained in the record, the place and the events are given correctly. In particular, the information must match the date, place, person, and period of the record [14, 29, 30].

However, the above-mentioned approaches alone may not be sufficient to analyze the trustworthiness of digital records. Because the legislation and the information technologies used as a source for the formation of the records have brought the issue to be discussed from a broader perspective. This perspective is called archival trustworthiness [14, 32, 33]. As with other notions of the trustworthiness, authenticity, accuracy, and reliability are critical [14, 32].

Authenticity, which is defined as the fact that the attributes of the record do not change during the period in which it is processed, filed, and archived after it has produced, is examined in two steps, identity and integrity. Identification refers to the qualification of the characteristic elements that distinguish them from other records and occurred according to their type. Examples of these are persons in the record, date of creation and transmission, subject, archival bond, file code, and appendix of the record. Another level of authenticity is integrity, which means that the record is undecomposed and unaltered, with all its components. It is aimed to preserve the context, form features and content of the record in integrity [14, 29, 34, 35].

In addition to authenticity, another element of trustworthiness is accuracy. An accurate record seeks to be precise, correct, consistent, and free from falsification. Reliability, which is another element of trustworthiness, is evaluated based on the completeness of the record form through the controls in the record production procedures. These controls are specified as the production and receiving of the record, its placement in its folder,

and the authorization of the persons in the records. The completeness of the record form refers to the presence of all elements of the intellectual form that make the record suitable for legal consequences [14, 29, 35]. Therefore, it was thought that the functions in the LoP could be related to the trustworthiness analysis developed by Sağlık. In this analysis, the trustworthiness of records is critiqued at the layers of records, technological conditions, organization, legislation, and society [14].

Figure 1. Layers of the Trustworthiness

The records layer evaluates the elements that make up the record such as context, archival bond, metadata and medium. Questions such as which metadata was used, whether or not a format change is required, and if the form elements were recorded are asked here. The technological conditions layer examines the application software and hardware used to produce, transfer, and store the records. Issues such as performing integrity checks, diversifying storage methods, and access privileges are analyzed. At the organization layer, policies and procedures regarding records management and archiving are evaluated. Issues such as the existence of a records management policy and the ongoing training of instructors are considered [14].

Although organizations prepare policies and procedures for records management and archiving, develop technological conditions in accordance with the needs of the service, and assign prospective metadata, they act in accordance with the relevant legislation while performing their functions. The legislation might include issues such as the retention period of the records, form elements, and technological conditions to be adopted. As such, these issues are critical elements of the legislation layer. Another aspect of this layer is evaluating the records management and archiving practices of the national archives. Therefore questions such as whether the national archives have determined the archiving rules for the records, whether migration procedures have been established, and the formats to be used have been specified [14].

The final layer of trustworthiness analysis examines what elements citizens look for to trust digital records. Therefore, this stage is called the society layer. Questions such as what are the tools that build trust among citizens, how much trust is placed in records, and how can this trust be increased will be explored [14].

Considering all these trustworthiness analyses, it is thought that institutions are more effective at records, technological conditions and organization layers. Because there are activities outside of the organizations' own savings at the layer of society and legislation. For example, at the community level, the opinions of citizens are critiqued, and at the legislation level, laws, regulations and circulars issued by government are reviewed. When this is the case, both citizen opinion and legislation are not directly in the hands of the organizations themselves. Organizations are more dynamic at the records, technological conditions, and organizations level [14].

It is possible to assess functions in the LoP according to trustworthiness layers. The functions may not be related to the same layers in all levels, for example, the first level of the Metadata function may be related to records layer, but in the second level it may be associated to the technological conditions. Table 1 shows the relation of LoP and the trustworthiness layer. R shows “Records”, T demonstrates “Technological Conditions” and the O indicates “Organization” layer.

Table 1. Relations of LoP to the Trustworthiness

Functions	**Level 1 (Know)**	**Level 2 (Protect)**	**Level 3 (Monitor)**	**Level 4 (Sustain)**
Storage	T, O	O	T, O	T, O
Integrity	R, T	R, T, O	R, O	R, T
Control	O	O	T	T, O
Metadata	R, T, O	R	O, R	R

Content	R	R, T	T	T

Since these functions are shaped by the activities of the organizations, the legislation and society layers of trustworthiness could not naturally find a place at the table. At the same time, a function may be related to the three layers in which organizations are more active. However, the layer is assumed to be formed directly by the corresponding function is indicated in the table. R shows records, T demonstrates technological conditions and O indicates organization layer.

The activities in the storage function are related to both to the technological conditions and to the organization layer. It is thought that, adopting a solid storage system is related to technological conditions; and keeping copies of the content in separate locations shapes the organization layer.

The integrity function is associated with almost all the trustworthiness layers. It is thought that generating integrity information structures the record layer, virus checking and backing up integrity information forms the technological conditions layer, and documenting integrity embodies the organization layer.

Determining access privileges in the control function is associated with the organization layer. Issues related to logs and audit trails are thought to shape the technological conditions layer. The metadata function is associated with almost all trustworthiness layers. It is thought that the creation of the inventory content is related to records, the backing up metadata is connected with technological conditions, and the determination of which metadata standards to apply is relevant to the organization layer.

Finally, the content function is associated with both the records and technological conditions layers. Identifying characteristics of the record embodies the records layer, and actions related to technological aspects of the record format such as emulation and migration figures the technological conditions layer.

IV. CONCLUSION

This study was an attempt to have a relationship between the LoP and the trustworthiness of digital records. Thus, it is intended to shed light on which layers of trustworthiness can be successful if organizations implement the functions in the LoP. The goals included in the LoP have been shown to be highly correlated with trustworthiness. These goals can be used as a benchmark when analyzing the trustworthiness of records created in organizations. The things that organizations should do to achieve the relevant goal can also be considered as trustworthiness criteria.

The LoP was developed to provide organizations with a goal in related functions. No mandatory criteria have been developed to allow flexibility for organizations. However, the lack of specific criteria in the LoP is considered a deficiency in terms of trustworthiness analysis. There is a need for criteria that are routinely checked and questioned for fulfillment.

As a result of the study, it has been seen that the trustworthiness of digital records can be successfully preserved after the realization of the LoP goals. However, examples of good practice are also in demand. This can be done by creating the criteria of the targets in the LoP. These criteria can be developed in a way that is flexible and not overly prescriptive.

1. REFERENCES

[1] Digital Preservation Coalition [DPC], “DPC Rapid Assessment Model”, *DPC*. [Online]. Available: https://www.dpconline.org/digipres/dpc-ram [Accessed: Mar. 06, 2023].

[2] Electronic Resource Preservation and Access Network [ERPANET], “ERPA studies ”, *ERPANET*. [Online]. Available: https://www.erpanet.org/studies/index.php [Accessed: Mar. 06, 2023].

[3] Cultural, Artistic and Scientific Knowledge for Preservation, Access and Retrieval [CASPAR], “CASPAR website”, *CASPAR*. [Online]. Available: http://casparpreserves.digitalpreserve.info [Accessed: Mar. 06, 2023].

[4] Preservation and Long-Term Access Through Networked Services [PLANETS], “Publications”, *PLANETS*. [Online]. Available: https://planets-project.eu/publications [Accessed: Mar. 06, 2023].

[5] Alliance Permanent Access to the Records of Science in Europe Network [APARSEN], APARSEN deliverables”, *APARSEN*. [Online]. Available: http://www.alliancepermanentaccess.org/index.php/about-aparsen/aparsen-deliverables [Accessed: Mar. 06, 2023].

[6] CoreTrustSeal, “Data repositories requirements”, *CoreTrustSeal*. [Online]. Available: https://www.coretrustseal.org/why-certification/requirements [Accessed: Mar. 06, 2023].

[7] Go FAIR, "FAIR principles", *GO FAIR*, [Online]. Available: https://www.go-fair.org/fair-principles [Accessed: Mar. 06, 2023].
[8] International Research on Permanent Authentic Records in Electronic Systems [INTERPARES], "INTERPARES Project", *INTERPARES*. [Online]. Available: http://www.interpares.org [Accessed: Mar. 06, 2023].
[9] B. M. Shabou, "Digital diplomatics and measurement of electronic public data qualities what lessons should be learned?", *Records Management Journal*, vol. 25, no. 1, pp. 56-77, Mar. 2015, doi: 10.1108/RMJ-01-2015-0006.
[10] R. D. Donaldson, "Development of a scale for measuring perceptions of trustworthiness for digitized archival documents", Ph.D. dissertation, University of Michigan, Michigan, United States, 2015. [Online]. Available: https://deepblue.lib.umich.edu/handle/2027.42/111489.
[11] R. D. Donaldson, "The digitized archival document trustworthiness scale", *International Journal of Digital Curation*, vol. 11, no. 1, pp. 252-270, Nov. 2016, doi: 10.2218/ijdc.v11i1.387.
[12] R. D. Donaldson, "Trust in archives–trust in digital archival content framework", *Archivaria*, no. 88, pp. 50-83, Nov. 2019. [Online]. Available: https://archivaria.ca/index.php/archivaria/article/view/13697.
[13] M. Ngoepe and J. Mukwevho, "Ensuring authenticity and reliability of digital records to support the audit process", INTERPARES, Jul 9, 2018. Accessed: Mar 6, 2023. [Online]. Available: http://interparestrust.org/assets/public/dissemination/AF06-FinalReport.pdf
[14] Ö. Sağlık, "Elektronik belge yönetimi uygulamalarındaki koşullar ışığında e-imzalı belgelerin delil değerinin arşivsel güvenilirlik açısından Incelenmesi", Ph.D. dissertation, Istanbul University, Istanbul, Türkiye, 2021. [Online]. Available: https://www.proquest.com/pqdtglobal/docview/2754923369.
[15] M. Schultz et al., "Building institutional capacity in digital preservation", in *Proc. 10th IPRES 2013*, J. Borbinha, M. Nelson, S. Knight, Eds. Sep. 2013, pp. 322-325.
[16] B. J. Daigle et al., "Level up on preservation: Updating and mapping the next generation of the Levels of Preservation", in *Proc. 16th IPRES 2019*, M. Ras, B. Sierman, A. Puggioni, Eds. Sep. 2019, pp. 512-513.
[17] M. Haunton, "Incorporating digital preservation and access maturity models into wider assessment programmes: Archive service accreditation and the levels of digital preservation and born-digital access", in *Proc. 18th IPRES 2022*, Sep. 2022, pp. 441-442.
[18] S. McMeekin, A. Currie, "Ain't no mountain high enough: Developing a new competency framework for digital preservation", in *Proc. 18th IPRES 2022*, Sep. 2022, pp. 99-107.
[19] M. Humbert, S. Roussel, E. Vasseur, "Building the future of digital preservation in French archival services: Processes, functions and staffing for an effective digital preservation", in *Proc. 16th IPRES 2019*, M. Ras, B. Sierman, A. Puggioni, Eds. Sep. 2019, pp. 46-52.
[20] P. Lucker et al., "Preservation watch at the National Archives of The Netherlands", in *Proc. 15th IPRES 2018*, Sep. 2018.
[21] J. van der Nat, M. Ras, "A Dutch approach in constructing a network of nationwide facilities for digital preservation together", in *Proc. 14th IPRES 2017*, Sep. 2017, pp. 99-107.
[22] F. Berghaus et al., "CERN services for long term data preservation", in *Proc. 13th IPRES 2016*, Oct. 2016, pp. 168-176.
[23] M. Pennock, P. Wheatley, P. May, "Sustainability assessments at the British Library: Formats, frameworks, & findings", in *Proc. 11th IPRES 2014*, S. Coates et al., Eds. Oct. 2014, pp. 141-148.
[24] NDSA. "Levels of Digital Preservation", *NDSA*. [Online]. Available: https://osf.io/QGZ98 [Accessed: Mar. 07, 2023].
[25] DigCurV. "DigCurV Curriculum Framework", *DigCurV*. [Online]. Available: https://digcurv.gla.ac.uk [Accessed: Mar. 07, 2023].
[26] DPCMM. "Digital preservation capability maturity model", *DPCMM*. [Online]. Available: https://www.securelyrooted.com/dpcmm [Accessed: Mar. 07, 2023].
[27] Ö. Külcü, "INTERPARES 3 kurumsal bilgi sistemleri içerisinde belge yönetimi: Türkiye'deki kamu üniversitelerinde gerçekleştirilen uygulamalara yönelik bir durum analizi", The Scientific and Technological Research Council of Türkiye, 2014.
[28] Ö. Sağlık, "Sayısal Koruma Koalisyonu Hızlı Değerlendirme Modeli: Elektronik belgelerin güvenilirliği açısından bir inceleme", *Bilgi Yönetimi*, vol. 5, no. 2, pp. 211-223, Dec. 2022, doi: 10.33721/by.1199232.
[29] Ö. Sağlık, "Arşivlenen elektronik belgelerin güvenilirliğini tehdit eden riskler: Teknolojik koşullar açısından bir inceleme. *Bilgi ve Belge Araştırmaları Dergisi*, no. 16, pp. 29-47.
[30] N. Çiçek, *Modern belgelerin diplomatiği*. Istanbul, Türkiye: Derlem Yayınları, 2009.
[31] H. MacNeil, *Trusting records: Legal, historical and diplomatic perspectives*. Springer, 2000.
[32] L. Duranti and R. Preston, Eds. *INTERPARES 2: Experiential, Interactive and Dynamic Records*. 2008. [Online]. Available: http://www.interpares.org/ip2/display_file.cfm?doc=ip2_book_complete.pdf. Accessed: Mar 7, 2023.
[33] J. Bushey, "The archival trustworthiness of digital photographs in social media platforms", Ph.D. dissertation, University of British Columbia, Vancouver, Canada, 2016. [Online]. Available: http://hdl.handle.net/2429/57606.
[34] C. Rogers, "Virtual authenticity: Authenticity of digital records from theory to practice". Ph.D. dissertation, University of British Columbia, Vancouver, Canada, 2015. [Online]. Available: http://hdl.handle.net/2429/52722.
[35] N. Çiçek, Ö. Sağlık. "Blokzincir teknolojisinin elektronik belgelerin güvenilirliğinin korunmasında başarıya katkısı, in *Bilgi Yönetimi ve Bilgi Güvenliği: eBelge-eArşiv-eDevlet-Bulut Bilişim-Büyük Veri-Yapay Zekâ*, B. Yalçınkaya et al., Ed., Ankara: Ankara Üniversitesi, 2019, pp. 141-170.

VIRTUALIZATION FOR PROCESSING AND ACCESSING DIGITAL ARCHIVES

Shelly Black
NC State University Libraries
USA
syblack@ncsu.edu
0000-0002-9046-4866

Brian Dietz
NC State University Libraries
USA
bjdietz@ncsu.edu
0000-0001-7190-2755

[Matthew] Farrell
Duke University Libraries
USA
matthew.j.farrell@duke.edu
0000-0003-1502-2651

Abstract – At a basic level, virtualization [1] is the use of a host computer or server's resources to run other computing environments. There are many ways in which virtualized computing environments may be deployed and interacted with, including using software to virtualize additional desktops on a local computer (e.g., VirtualBox, Hyper-V Manager, or VMWare) or accessing virtual command line interfaces hosted by a server or computer cluster [2],[3], and emulating old video game systems on contemporary hardware [4]. In this paper we discuss a cross-institutional collaboration on using containerization and desktop virtualization in digital curation at academic special collections libraries.

Keywords – Containerization, desktop virtualization, virtual machine, special collections, born-digital archives, virtual reading room

Conference Topics – We're All in this Together; Sustainability: Real and Imagined

I. INTRODUCTION

While there are many applications for virtual machines, we highlight two general affordances. First, a user accessing a virtual machine can make use of software and processes in a different operating system, such as Linux-only toolsets on a Windows host. Second, virtual machines allow users to use specific or unique computing environments remotely. Applying these specifically to born-digital special collections work, running virtualized environments allow staff and researchers to access consistent toolsets and configurations regardless of the host computer(s) in use. The authors' staff computing environments are composed of multiple, varied physical configurations, but through the use of virtual environments, each of our workstations can make use of consistently packaged, identical processing environments for technical services workflows. In terms of public services work, virtualized environments allow one to create controlled environments for researchers to access digital archival materials remotely.

II. CONTAINERIZATION FOR PROCESSING

Container technology allows one to package applications and dependencies into a Linux environment so that they can be tested and deployed and trusted to work consistently across computing platforms [5]. Compared to other types of virtualization, containers are usually defined to contain only the resources necessary to complete a specific set of tasks as opposed to an entire operating system. Separate containerized applications or "services" can be run together in an orchestrated way. For instance, an application may include separate containers for a web service, database, and SOLR index. Additionally, invoking a process in one container may call additional containers to perform additional automated or semi-automated processes. Docker and Podman are two popular platforms that support containerization.

For several years, North Carolina State University Libraries (NC State) has managed the majority of its born-digital processing tools using Homebrew for Mac [6], along with pip for installing Python packages. Duke University Libraries has managed Windows computers that run the BitCurator environment as a virtual machine and in a dual boot configuration. Both organizations were motivated to find a more lightweight and flexible approach to managing applications, and one that might avoid the complications of updating working environments following new releases and installing tools on new machines.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

In 2020, NC State started to examine container technology to address these issues [7]. Doing so would simplify installation and management of command line tools; better support cross-platform replication, functionality and user experience; and result in a shareable and replicable approach. Duke joined the process as a collaborator in 2021.

Early explorations began by defining a minimal viable product (MVP): a container one could use to perform virus scans, search for personally identifiable information (PII), and conduct file format characterization on files accessible via the host computer. NC State initially attempted to create an image using an official Docker build of Linux Homebrew [8], drawing on past experience working with Homebrew for Mac. When this proved infeasible, the next attempt was to build an image based on the official Docker build of Kali Linux, installing their "forensic metapackage" of applications [9]. This reached MVP, but it was ultimately decided that the extent of tools available in the metapackage resulted in a bloated image and container. Drawing on this success, we focused on using official Docker builds of Ubuntu and Fedora Linux [10], which resulted in the creation of a more tightly scoped image, i.e., one that excludes extraneous tools. Recent testing coincided with both organizations purchasing or assessing Apple computers with the Apple Silicon ARM chips, leading to the creation of containers based on the ARM Linux image. This period of iteration confirmed an early assumption: that adding to or otherwise updating a container is more efficient than performing similar maintenance across multiple standalone workstations.

To date, NC State and Duke have written Dockerfiles that contain instructions for building an Ubuntu-based AMD64 Linux image and Fedora-based AMD64 and ARM Linux images [10]. The container used at each institution during processing is derived from these images. At both NC State and Duke, the container environment includes command line applications for searching for sensitive data and duplicates, virus and malware detection, and file characterization, as well as general Linux file utilities. With these toolsets, the containers support the same range of files and content types as is currently supported in systems such as the BitCurator environment. However, some steps in our workflows will continue to be done on the host. Disks can be shared as volumes, and their files can be packaged from within the container. Yet containers cannot access devices, unless used on a host Linux computer. Disk imaging and optical disc audio ripping must be completed on host Mac and Windows workstations. Our containers are currently deployed on these three host operating systems, and we are using containers in production or expect to be by fall 2023.

III. Desktop Virtualization for Access

NC State uses desktop virtualization to provision a remote virtual reading room service. A special collections reading room is traditionally a mediated environment where researchers can use materials. In some cases, there are copyright, privacy, or other donor-imposed access restrictions [11]. This applies to physical and digital materials. In the past, researchers at NC State who requested born-digital or digitized materials had to use an air gapped [12] laptop in the reading room. Specifically, WiFi and USB device access were disabled, so that researchers could not transfer the materials to themselves. Desktop virtualization allows NC State to replicate this secure environment for accessing digital materials online, eliminating the need for travel, and allowing multiple researchers to use it simultaneously.

Some institutions use digital asset management systems (DAMS) which function as virtual reading rooms [11, pp. 162-163]. These are appropriate for materials that can be openly shared, and for file formats that can be rendered in a browser or downloaded for viewing locally. However, maintaining a DAMS can be labor intensive. NC State's virtual reading room relies on existing infrastructure provided by the university's Virtual Computing Lab (VCL) [2]. This on-demand, virtualized computing service allows classes and researchers to connect to a remote server using Remote Desktop Protocol [13] software and access custom software environments.

In 2020 NC State began working with VCL on our server reservation. We created an Ubuntu Linux image that contains software and networking configurations from which the server can reboot. We installed open source software for viewing text documents, images, videos, and other common file formats, as well as a module to redirect sound.

Security configurations include firewall rules blocking HTTPS traffic, a disabled SSH client, and disabled drive and clipboard redirection. Thus, researchers cannot copy, download, or email materials to themselves. Linux permissions are also applied, so that the researcher can only view the files they requested. We cannot prevent them from taking screenshots. However, when they request to use the virtual reading room, they agree that materials are non-circulating and any pictures taken are for research purposes only. Another safeguard is that the virtual reading room can only be accessed with NC State credentials or by external researchers who create accounts with VCL. Administrative access to the server is controlled by an access group, to which staff were added through the VCL website.

The virtual reading room is currently an active service, having been used by five researchers in the 2022-2023 academic year. All researchers have succeeded in accessing and viewing their desired materials, with one exception, where the researcher could not connect for unidentified reasons. Feedback provided by researchers has been encouraging. We also receive regular inquiries from other institutions on how to implement this service, and Duke is interested in exploring or adapting NC State's approach for use with its patrons.

IV. Lessons Learned and Future Work

Once we have more production experience with containers, there are additional areas of exploration to consider. This includes best practices in building images and efficiencies in maintaining them. While we currently use one container for all processes, we may further explore whether and when to split our containers into separate, coordinated, specialized services, such as those for processing email archive files or used in post-processing work. We are also eager to explore the extent to which containers might support certain automated workflows.

Testing is also anticipated for the virtual reading room. It is most likely that researchers would request text documents, images, videos, or other common file formats. However, future use cases may include providing access to less common file formats, such as those used in computer-aided design, or an emulator to run legacy software. Overall, the user experience for researchers can be improved. This includes video streaming quality when using Microsoft Remote Desktop for Mac. Using assistive technology with the virtual reading room also needs to be tested. Additionally, ongoing maintenance involves ensuring that the virtual reading room uses a currently supported version of Ubuntu. It was originally installed with Ubuntu 18.04 LTS, which having just reached its end of life [14], required an upgrade. Because the security configurations were scripted, setting up a Ubuntu 22.04 LTS server as the virtual reading room required little effort.

Setting up, maintaining, and using both containers and virtualized desktop environments requires some degree of technical knowledge. As we deploy containers into full production, we will be gathering feedback from full-time and student staff, particularly to better understand gaps in technical skills. To use the environments, they need a basic working knowledge of a shell and the Linux file system. More technical knowledge is required to administer containers and customized virtual desktops, including a general understanding of virtual computing. Specific knowledge is required for building, deploying, updating, and managing these environments. That said, the authors are self-taught and do not have formal backgrounds in systems administration or IT desktop support.

The projects in this presentation started at a single institution before expanding to a peer organization [15], but wider distribution has been a consideration since the earliest stages. Our containerized processing environments can easily be distributed via Git as Dockerfiles, and can be reused, amended, and otherwise modified from the base versions to fit the use cases of other institutions. Similarly, the shell script to configure the virtual reading room can be shared and applied to a virtualized desktop hosted by other institutions or cloud computing services. We believe virtualization can increase the availability of processing environments and digital special collections for staff and researchers, respectively.

1. REFERENCES

[1] IBM. "What is Virtualization?" https://www.ibm.com/topics/virtualization (accessed Mar. 1, 2023).

[2] NC State University. "Virtual Computing Lab." https://vcl.ncsu.edu/ (accessed Mar. 1, 2023).

[3] Duke University. "Virtual Computing Manager." https://vcm.duke.edu/ (accessed Mar. 1, 2023).

[4] Rhizome. “Rhizome to Restore and Present Theresa Duncan CD-ROMs.” https://rhizome.org/editorial/2014/nov/18/announcing-theresa-duncan/ (accessed Mar. 1, 2023).

[5] A. Gaitonde. “Introduction to Containers: Basics of Containerization.” https://medium.com/geekculture/introduction-to-containers-basics-of-containerization-bb60503df931 (accessed Mar. 7, 2023).

[6] Homebrew. https://brew.sh/ (accessed Mar. 7, 2023); NCSU-Libraries. “bd-brewfile.” https://github.com/NCSU-Libraries/bd-brewfile (accessed Mar. 7, 2023).

[7] B. Dietz. (17 Nov. 2021). Lightweight Distribution of Tools. Presented at 2021 BitCurator Users Forum. [Online]. Available: https://docs.google.com/presentation/d/1-y2tVJc6TOsV4Ahb-gAJaAAJKxje-obHIb7ZnQW6P4I/edit#slide=id.gf344f557f2_2_50 (accessed Jun. 22, 2023).

[8] Docker Hub, “Homebrew.” https://hub.docker.com/u/homebrew (accessed Jun. 26, 2023).

[9] Kali Linux, “Containers.” https://www.kali.org/get-kali/#kali-containers (accessed Jun. 22, 2023); Docker Hub, “Kali Linux.” https://hub.docker.com/u/kalilinux (accessed Jun. 26, 2023); Kali-Meta, “kali-tools-forensics.” https://www.kali.org/tools/kali-meta/#kali-tools-forensics (accessed Jun. 26, 2023).

[10] Docker Hub, “Ubuntu,” https://hub.docker.com/_/ubuntu. (accessed Jun. 26, 2023); Docker Hub, “Fedora.” https://hub.docker.com/_/fedora (accessed Jun. 26, 2023).

[11] E. Arroyo-Ramírez, et al., “Speeding Towards Remote Access: Developing Shared Recommendations for Virtual Reading Rooms,” in *The Lighting the Way Handbook: Case Studies, Guidelines, and Emergent Futures for Archival Discovery and Delivery.* Stanford, CA: Stanford University Libraries, 2021, pp. 141-167. [Online]. Available: https://doi.org/10.25740/gg453cv6438 (accessed Mar. 7, 2023).

[12] National Institute of Standards and Technology, “air gap - Glossary.” https://csrc.nist.gov/glossary/term/air_gap (accessed Jun. 27, 2023)

[13] Microsoft, “Understanding the Remote Desktop Protocol (RDP),” https://learn.microsoft.com/en-us/troubleshoot/windows-server/remote/understanding-remote-desktop-protocol (accessed Jun. 27, 2023).

[14] L. Sandecki. (14 Mar. 2023). “Time to prepare for Ubuntu 18.04 LTS End of Standard Support on 31 May 2023.” https://ubuntu.com//blog/18-04-end-of-standard-support (accessed Jun. 30, 2023).

[15] S. Black, B. Dietz, and Farrell. (12 Jul. 2022). Virtual Computing for Digital Special Collections. Presented at 2022 Triangle Research Libraries Network (TRLN) Annual Meeting. [Online]. Available: https://docs.google.com/presentation/d/1Lxka4dweKIBON3ctx7zzbrfsO0yVGMP7WKAPW-59PKk (accessed Jun. 28, 2023)

MONITORING FILE FORMAT OBSOLESCENCE IN REPOSITORIES

An applied method

Sam Alloing
National Library of the Netherlands
Netherlands
sam.alloing@kb.nl
0000-0002-1254-1483

Abstract – The Dutch Digital Heritage Network (DDHN) wants to improve the monitoring of file format obsolescence. The Preservation Watch group researched on how institutions can monitor the life cycle of file formats in their repositories and how the monitoring could be implemented on a broader scale. Monitoring file format life cycle implies there needs to be a way to measure format obsolescence or helps an institution to identify when a file format is getting obsolete. The applied research identified the needed information and used a known model to search for trends and is applied in widespread areas. The model was compared with a naive method to evaluate the more complex method. This approach was tested in different types of repositories and used different file formats to research the robustness of the approach. This paper will investigate the possibilities and shortcomings of this method and further research that is required.

Keywords – preservation watch, file formats, applied research, file format obsolescence, Bass diffusion model

Conference Topics – From Theory to Practice

I. INTRODUCTION

Format obsolescence is a widely discussed topic in digital preservation. There are different strategies dealing with obsolete file formats like file format migration or emulation. The moment when to execute the preservation strategy is not an easy decision. Some policies use a late migration strategy. This strategy needs information about when to take a preservation action, so the migration is not too late and files can still be opened.

This paper uses the outcome of an earlier paper [1] that investigated the Bass Diffusion Model as a possible solution for detecting file format obsolescence and builds upon the results by using repositories of different institutions. The Bass Diffusion Model is used in a wide variety of use cases and is not specific for digital preservation.

II. METHODOLOGY

The increase and decline of products is described and predicted in the Bass Diffusion Model [2]. The model describes the life cycle of a product where innovators are early adopters of a product and later the imitators join with the big increase in use and the diminishing effect of laggers that follow after that. This gives the curve a typical bell shape with a steep start and a long tail [3]. Depending on where a product is in its life cycle it shows a cut out of the bell shape.

This model was also previously applied in the area of file format obsolescence and deemed useful. The model was applied in a context of a web archive and this has limitations on which file formats can be researched. The repositories of an institution are also different then the corpus of a web archive, because the last one has predominantly file formats that are used on the web, like for example HTML[4].

To help the interpretation of the output of the Bass Diffusion model a second model was used as a reference model. The linear regression model is used, because it is a simple model that represents the naive approach. To be useful the more complex model needs to be a better explanation then the

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

simpler model otherwise there is no added value. Because the simplicity of the linear regression, it is only applied on file formats with declining popularity. This way linear regression can be used as an evaluation model.

The aim of the research is also to look into the prediction capabilities of the models and we use the last quarters as a test set for prediction. In the plots it is shown as a green (Bass test) and purple line (Lineair test). Most of the data is used as a training set and the smaller test set the model needs to predict the course of the life cycle of the file format. This is used as an indication of the reliability of the prediction by the different models. The Blue line on the plot is the number of files.

III. Used approach

The approach looks for diminishing delivery to an institution or use on the internet of a file format over time. This is an indication used in the model as a diminishing popularity of a file format. The time period is over several years. Because there are also rare file formats the time period is over several quarters in a year so there are enough data points to make a predicting model.

The life cycle can include an increase and a decrease of popularity and shows at which stage in the life cycle a file format is. This also brings up the question if there is a threshold which indicates if a format is getting obsolete or if the file format monitoring can be automated. Not only the monitoring of a single file format is investigated, but also if file formats are linked together and if a file format is a predecessor which shows a decrease in popularity and if there is a successor that shows an increase in popularity .

A last and final factor that is important to monitor is the relation between applications and file formats. A decrease in the number of applications that can open or write a certain file format over time gives an indication of a file format becoming obsolete, because a decrease in popularity of a file format doesn't need to mean obsolescence. The combination of file formats and applications will be used as an indication of obsolescence.

IV. Data quality

For the analysis we used two types of data, data from Common Crawl and data from different institutional repositories (Netherlands Institute for Sound and Vision and Data Archiving and Networked Services (DANS)). The Common Crawl data is publicly available data from internet crawls. There are summaries available of for example mime type and this prevents the need to process all the Common Crawl dataset [5]. The mime types are identified by Apache Tika [6]. The summarized metadata was used for this analysis. The disadvantage is that only date from 2017 onwards was usable, but in general this is also the year in which Common Crawl data is more reliable [7]. This limits the results of the output as an indication of the file format lifecycle on the broader and international scale.

This is a recurring theme, getting usable information is a challenge also for repositories of institutions. Although the information seems simple, just the date that indicates the creation of the file or a substitute date of the resource that is preserved like publication date. But that was a challenge. Most institutions were able to produce ingest dates, but that isn't a sufficient date, for example because of migrations of content when changing systems. The other challenge was to produce file format identification that was precise enough. Most institutions could only produce mime types or file extensions which don't have file format version information. For example the MS Access database format, MDB, can contain a wide range of MS Access database software versions [8]. Institutions using for example the Pronom PUID which can describe the MS Access file format much more precise [9].

V. Analysis and results

A. Common Crawl

The Common Crawl analysis was used to test the approach on a large scale data set. The hypothesis is to use this as a comparison to the repository level and use this as an extra evaluation criteria to interpret the results of the repository level.

Of the different analyzed mime types XHTML and GIF will be discussed [10].

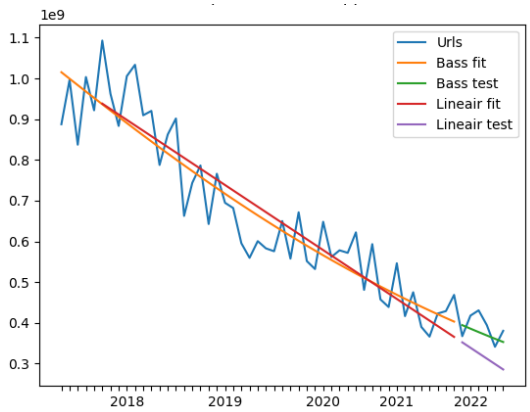


Fig 1 XHTML plot. 2022 CC-BY-SA-4.0 Rein van 't Veer/DDHN

The XHTML plot (Fig 1) shows a declining graph. The format is in the downwards spiral of the bell shaped curve, but is not in danger of getting obsolete. The format still constitutes 12% of the billions of pages harvested by Common Crawl, so no obsolescence is expected. Browsers can still open the file format as well.

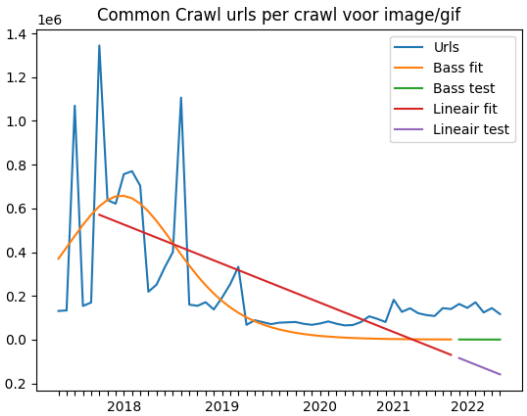


Fig 2 GIF plot. 2022 CC-BY-SA-4.0 Rein van 't Veer/DDHN

The GIF file format (Fig 2) is also in a decline, but there is a part that shows an increase. The increase is due to the incomplete set. The file format is already in use for a long time, but the limited time period of the data set and the erratic peaks throw off the Bass model. The linear regression just shows a decline, but the prediction goes below 0. The Bass model shows a more realistic trajectory.

Of the 26 investigated formats in the Common Crawl data set, the Bass model had in 13 cases a better prediction than the linear regression. In 3 cases the accurateness was the same and in 8 cases the linear regression performed better. The reason for these errors is probably comparable to the GIF case already discussed, the erratic peaks. This is also suggested by the other plots of other data sets.

B. Data Archiving and Networked Services (DANS)

DANS [11] is an institute in the Netherlands that preserves scientific data from scientific institutes. The data set from the archaeological repository is used. In this data set the analysis of multiple linked file formats was possible. The Microsoft Office formats (MS Word and Excel) show a linked file format lifecycle between different formats. The case of MS Excel formats XLS (Fig 3) and XLSX (Fig 4) is discussed as it shows the evolution clearly.

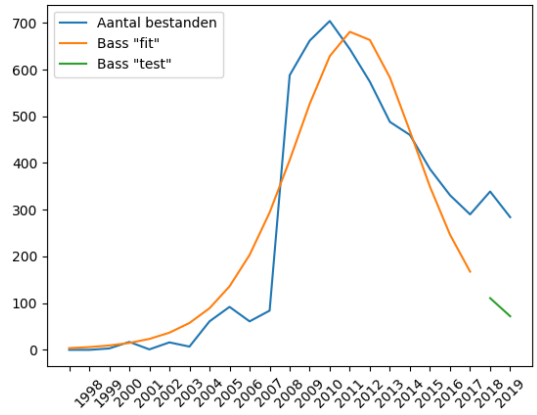


Fig 3 XLS plot. 2022 CC-BY-SA-4.0 Rein van 't Veer/DDHN

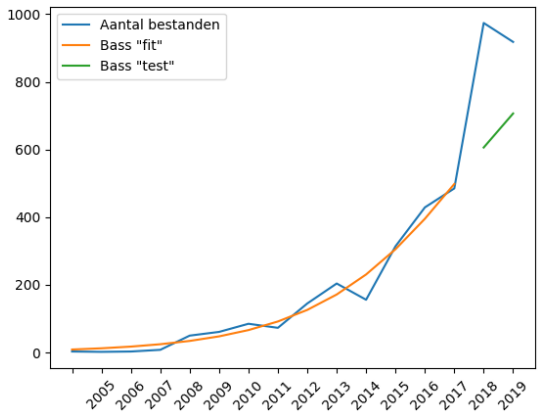


Fig 4 XLSX plot. 2022 CC-BY-SA-4.0 Rein van 't Veer/DDHN

These two plots show the decline of the XLS format and around the same time an increase of the XLSX format. This is expected as the XLS format is an older format that has been gradually phased out by Microsoft in favor of XLSX [12]. Microsoft Access doesn't show this trend in the DANS repository, the MDB file format (Fig 5) which is older than the ACCDB file format (Fig 6) still is very popular and is still increasing. This is unexpected, but can be explained by the specific case of archaeological data sets where MS Access is popular software and database templates in MDB file format are used and reused over and over. Also the number of files received is

much lower for the ACCDB file format (Blue line). This throws off the Bass Model prediction with a sharp increase in ACCDB and a decrease in the case of MDB.

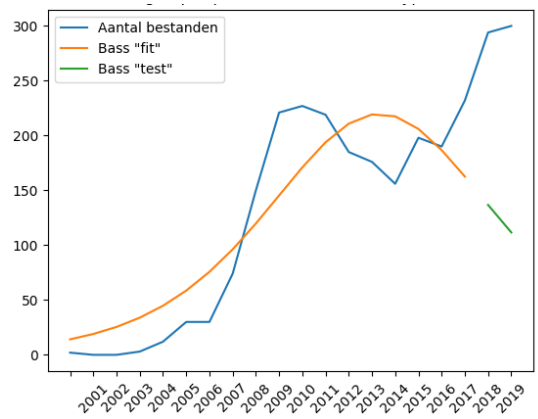


Fig 5 MDB plot. 2022 CC-BY-SA-4.0 Rein van 't Veer/DDHN

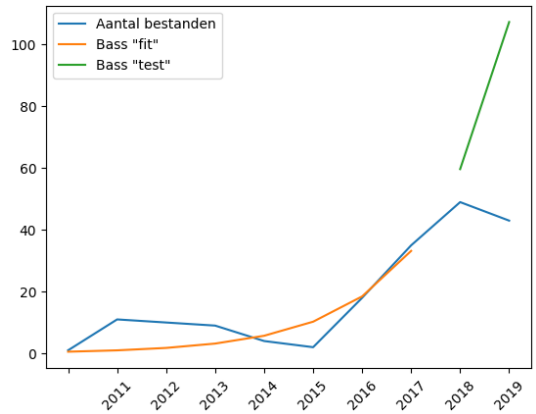


Fig 6 ACCDB plot. 2022 CC-BY-SA-4.0 Rein van 't Veer/DDHN

The DANS data set contains more file formats, these are not discussed here, but are described in an article [13].

VI. Application as extra dimension

It becomes apparent by the DANS data that file format is not the only dimension to look at when analyzing the file format life cycle. There is a need for an extra dimension, what application still supports the file format.

To start this analysis a good source of information is necessary. Different possible sources have been researched: Guide of preferred file formats [14], NARA digital preservation framework [15], Wikidata [16] and Pronom [17]. To evaluate the sources the case of Microsoft Access was used. During the analysis of the results of the DANS data set the MDB file format showed declining support in Microsoft Access and is in danger of getting obsolete [18]. This case shows that there is fine grained information needed between file format and application. The data model of Pronom and Wikidata can store the information that is needed to support the research. The problem with Pronom is the information not kept up-to-date [19]. The Wikidata data model has the potential to support the connection between file format and application, but the link is not yet sufficiently provided. The application version information is a literal and not an entity. A literal is a string of information and is not easy to query or it is not possible to link information to a literal. This is all possible with an entity, but in the case of Microsoft Access, this is most of the time not available. For example MS Access file format version 95, has as software version identifier 95 [20]. Microsoft Access Database, version 2007 [21] is an entity and queries are possible of for example the number of applications that can read the file format [22]. This shows potential but needs to be researched more and more data needs to be added like for example discontinued date [23].

VII. Conclusion

The research shows that the Bass Model can be used as a method to evaluate the format obsolescence, but it is not an automated process because the results need to be interpreted and understood in the specific context of a repository or in the broader scale, due to the specific community the repository serves or due to data quality issues. The method helps with summarizing the file format information and gives insight in the life cycle of the file format. The relation between the broad internet scale data set and the repository level data sets needs more research because of limited data sets and different file formats researched.

The relation between file format and application needs to be researched more, certainly if the analysis needs to be combined with the file format information and help to improve the file format life cycle analysis.

VIII. Acknowledgement

This research was conducted by Rein van 't Veer of Antfield Creations and The Preservation Watch working group of The Dutch Digital Heritage Network and.

The Dutch Digital Heritage Network is formed by organizations in the fields of culture, heritage, education, and research together. With suppliers of heritage software, provinces and municipalities we are working on the implementation of the National Strategy Digital Heritage, supported by the Ministry of Education, Culture and Science.

REFERENCES

[1] Duretec, K. and Becker, C. (2017), Format technology lifecycle analysis. Journal of the Association for Information Science and Technology, 68: 2484-2500. https://doi.org/10.1002/asi.23881

[2] Bass, Frank M. (1969), A New Product Growth for Model Consumer Durables, Management Science, Vol. 15, No. 5, Theory Series, p. 215-227 https://math.la.asu.edu/~dieter/courses/APM_598/Bass_69.pdf

[3] Bass diffusion model. https://en.wikipedia.org/wiki/Bass_diffusion_model

[4] See HTML in MIME Types. https://commoncrawl.github.io/cc-crawl-statistics/plots/mimetypes

[5] MIME Types. https://commoncrawl.github.io/cc-crawl-statistics/plots/mimetypes

[6] Apache Tika - a content analysis toolkit. https://tika.apache.org/

[7] Size of Common Crawl Monthly Archives. https://commoncrawl.github.io/cc-crawl-statistics/plots/crawlsize

[8] Microsoft Access MDB File Format Family. https://www.loc.gov/preservation/digital/formats/fdd/fdd000462.shtml

[9] Wikidata Query about file format with file extension MDB and PUID https://w.wiki/6QXs

[10] For more mime types see (Dutch only): Rein van 't Veer (2022) Monitoring van bestandsformaten 2: het internet als archief, het Bass-model in de praktijk: welke internetformaten zijn aan het verdwijnen? https://kia.pleio.nl/groups/view/4fc4e83a-f55b-4000-b1cb-3fe9a16d3f93/kennisplatform-preservation/blog/view/b4b724a2-0683-438d-897c-717b95a57071/monitoring-van-bestandsformaten-het-internet-als-archief-het-bass-model-in-de-praktijk-welke-internetformaten-zijn-aan-het-verdwijnen

[11] DANS | Centre of expertise & repository for research data. https://dans.knaw.nl/en/

[12] Library of Congress, XLSX Transitional (Office Open XML), ISO 29500:2008-2016, ECMA-376, Editions 1-5 https://www.loc.gov/preservation/digital/formats/fdd/fdd000398.shtml, see Adoption

[13] Rein van 't Veer (2022) Monitoring van bestandsformaten 4: formaten in gebruik bij Data Archiving and Networked Services (DANS) https://kia.pleio.nl/groups/view/4fc4e83a-f55b-4000-b1cb-3fe9a16d3f93/kennisplatform-preservation/blog/view/b92f7a1e-fd6a-4c88-875c-5bccc470554c/monitoring-van-bestandsformaten-formaten-in-gebruik-bij-data-archiving-and-networked-services-dans (Dutch only)

[14] DDHN, Summary Guide to Prefered Formats. https://www.wegwijzervoorkeursformaten.nl/index.php/Summary_Guide_to_Prefered_Formats

[15] NARA, U.S. National Archives and Records Administration Digital Preservation Framework. https://github.com/usnationalarchives/digital-preservation#the-nara-risk-and-prioritization-matrix

[16] Wikidata. https://www.wikidata.org/wiki/Wikidata:Main_Page

[17] PRONOM https://www.nationalarchives.gov.uk/PRONOM/

[18] Microsoft, Which Access file format should I use? https://support.microsoft.com/en-us/office/which-access-file-format-should-i-use-012d9ab3-d14c-479e-b617-be66f9070b41

[19] This is confirmed by Francesca Mackenzie, Digital Archivist at UK National Archives and responsible for Pronom.

[20] Wikidata, Microsoft Access Database, version 95. https://www.wikidata.org/wiki/Q48004869

[21] Wikidata, Microsoft Office Access 2007. https://www.wikidata.org/wiki/Q46049725

[22] Wikidata query, Applications that can read MS Access file formats https://w.wiki/6RMa

[23] Wikidata, discontinued date. https://www.wikidata.org/wiki/Property:P2669

COMMUNITY ARCHIVES AT THE DIGITAL REPOSITORY OF IRELAND

Lisa Griffith
Digital Repository of Ireland
Ireland
l.griffith@ria.ie
0009-0007-3651-1477

Kevin Long
Digital Repository of Ireland
Ireland
k.long@ria.ie
0000-0002-1041-2661

Abstract – The Community Archive Scheme is a bottom-up method of community engagement that the Digital Repository of Ireland (DRI) uses to work directly with no or low-income groups with digital material to preserve. The DRI's usual depositors are academic, cultural heritage, or public organisations and libraries with a long history of archiving who select material from their own collections for preservation. Through the Community Archive Scheme, we work in a hands-on way to provide digital preservation to a wider range of groups that fall outside of this sphere. The scheme celebrates its fifth anniversary in 2023 and during this period DRI has worked with nine voluntary groups to help preserve material on a variety of topics including the experience of asylum seekers in Ireland, maternal health, built heritage, LGBT rights and activism in Ireland. The types of material that we are working to preserve through this scheme vary from photographs of artists' works such as quilts, audio-visual material such as community documentaries, and documentaries produced for digital radio and social media. This paper will discuss how the scheme evolved, how these organisations have strengthened DRI as an organisation as well as making our community and collections more equitable and diverse, challenges we have encountered, some of the solutions we have developed, where our successes have come from and some of the future developments we are exploring so that we can continue to work with these groups.

Keywords – Digital Archives, Community Archives, Digital Preservation, Inclusion, Membership, Ireland, Cultural Heritage data

Conference Topics – DIGITAL ACCESSIBILITY, INCLUSION, AND DIVERSITY; WE'RE ALL IN THIS TOGETHER

1. INTRODUCTION

DRI is a research-performing organization and national Trustworthy Digital Repository (TDR) for Ireland's humanities, cultural heritage, and social sciences data. DRI has been certified by the CoreTrustSeal since 2018. As a national infrastructure for the arts, social sciences, and humanities, DRI provides reliable, long-term, sustained access to social and cultural digital data. We make this data openly available in line with the FAIR data principles of findability, accessibility, interoperability, and reusability. We aim to safeguard Ireland's social, cultural, and historical record through active management of digital content over time to ensure that this content remains accessible to researchers, cultural heritage enthusiasts, and members of the public into the future. We support best practices in digital archiving, digital preservation, Open Access, Open Research, and FAIR data sharing. DRI is funded by the Department of Further and Higher Education, Research, Innovation and Science (DFHERIS) via the Higher Education Authority (HEA) and the Irish Research Council (IRC).

The route for adding collections to DRI is through paid membership. We have forty paid members from Ireland's higher-level institutions, local authorities, research groups and centres, galleries, libraries, archives and museums. It is important to say that DRI does not take ownership of the collections that are published on the Repository, we steward them. Copyright remains with the collection owners. Where they arise, questions, or decisions, about the use of these collections are sent to the collection owners. The terms and conditions that govern the management of these collections are laid down in our Organizational Manager Agreement.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

Our federated model of membership means that the collection owners are given access to the repository so that they can ingest the collections and the accompanying themselves. The benefits we offer to members include ingest to the Repository, long-term preservation of collections, training on digital preservation, training on how to add your collection to the repository, and access to the advice of DRI staff and our events. Digital objects published on the Repository are issued with DataCite DOIs. This paid-membership scheme was launched in 2018 and DRI has grown by approximately eight paying members yearly. The paid membership scheme, offering full membership at €5,000 per annum or associate membership at €500 per annum, was launched in March 2018.

2. The Beginnings of the Community Archive Scheme

As we were preparing to launch the paid membership scheme we began to discuss the types of groups that would be excluded because of the cost of membership and how we might create a different route for them to access digital preservation in the repository. Over the next six months, we discussed how we might offer the benefits of membership to low or no-income groups and who those groups might be. DRI was still in the stage of digital preservation education for Irish HSS and Cultural Heritage audiences, so we didn't know who might apply to the scheme, what level of support they would need, or what the size of their collections might be. We just wanted to make sure there was a space within our organization for groups who might find our membership fees prohibitive. Our Collection Policy also mandates a focus on at-risk data and topics underrepresented in the repository, and we suspected a convergence between groups holding this material and those who could not afford membership.[1] With that in mind, the conditions for the scheme were that 'no or low-income groups' could apply and would receive associate member benefits for a year. The scheme was launched in late 2018 for the following year.

Criteria for eligibility are that the materials organizations are seeking to deposit are already a digital format and that they have volunteers in their organization who have time to attend training and deposit the collections. We also ask that they have metadata to go with the collections (though we provide assistance in meeting our minimum system requirements). Finally, we ask that they have copyright clearance to deposit the items in their collections.[2] In the first year, it was clear that there was a demand for the scheme when we received 8 applications. The inaugural winner of the award was the Cork LGBT Archive which was run by Orla Egan, an activist who had a strong archival focus. The collection was well organized, highly curated, had strong metadata, and Orla was very familiar with the work of DRI so she did not require much additional training. Her collection was published in 2019.[3] Two groups were awarded under the scheme in 2020, the Asylum Archive and Cork Media Framework, in recognition of the outreach challenges that Covid 19 brought to large and small organizations. We awarded three groups in 2021 including Joe Lee Films,[4] Dublin Ghost Signs and the Elephant Collective. In 2022 Tulsk History Society[5] and Bray Arts[6] were awarded under the scheme and in 2023 the winner was Dublin Digital Radio.

---

[1] Digital Repository of Ireland. (2021) DRI Collection Policy, Digital Repository of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.kk91v774c-2

[2] Digital Repository of Ireland, Community Archive Scheme https://dri.ie/dri-community-archive-scheme

[3] Orla Egan. (2019) Cork LGBT Archive, Digital Repository of Ireland [Distributor], Cork LGBT Archive [Depositing Institution], https://doi.org/10.7486/DRI.2j635q62d

[4] Joe Lee. (2021) Dublin based community films by Joe Lee, Digital Repository of Ireland [Distributor], Joe Lee Community Based Films [Depositing Institution], https://doi.org/10.7486/DRI.90205r016

[5] Tulsk History Society. (2022) Tulsk History Society: Letters from the 1880s - 1890s, Digital Repository of Ireland [Distributor], Tulsk History Society [Depositing Institution], https://doi.org/10.7486/DRI.7h14qf91p-1

[6] Bray Arts. (2022) Bray Arts Collections, Digital Repository of Ireland [Distributor], Bray Arts [Depositing Institution], https://doi.org/10.7486/DRI.5t356b38v

3. Programme Success

The immediate and visible success of the Community Archive is that it has increased the number and diversity of our datasets. Many of these collections intersect, thematically or geographically, with collections we already or subsequently held which means the datasets can take on new layers and meaning. The Community Archive Scheme has also brought us into contact with Repository users who have accessibility issues. It has become more common to have users with accessibility issues on DRI's website and the Repository interface but we have worked with these users to upload their collections. We have had to improve our site accessibility and have had the opportunity to work in a hands-on way with these users.

Working with these groups has given us an insight into just how vulnerable their material is. In addition to all of the usual threats like bit rot and digital obsolescence, the material is threatened because of a lack of funding or lack of appreciation for what information it holds. As much of the material lies outside formal organizations, it is held by volunteers with low, or most likely, no income. The material is often preserved by one custodian with an active interest and appreciation of its importance. If this custodian moves on, or cannot afford to sustain the material, it is lost. Working with these nine community groups has made sure that we are preserving endangered material and both the groups and the staff working on these collections value this work. Digital Radio Recordings, for example, have been described as 'endangered' in the Digital Preservation Coalition's 'Bit List' in 2022.[7] In a broader organizational sense, we have learned a lot from working with these groups and these lessons have fed into our research projects such as our Wellcome Trust-funded 'Archiving Reproductive Health' project.

4. Challenges and Solutions

In addition to the very clear benefits, the work presents several challenges for DRI. These challenges can be practical, technical, outreach or policy-related, or even organizational challenges. Applications to the scheme can be uneven and vary from year to year. Often we are approached by groups whose collections are interesting and at-risk, but not yet sufficiently digitized to be eligible for the scheme. While we run an active social media and targeting campaign, word of mouth is sometimes the best way to find suitable groups. This indicates though that there is a general lack of awareness about digital preservation, what it means and the processes that are needed to support it.

Associate membership was offered to community archives for a year. We now know that a year is not enough for many community archives to ingest their collections as most are unpaid volunteers. Unlike with the mainstay of our members which are organizations with libraries and archives, we have realized that we need to think about the technical language we have been using at our training sessions and in the supporting material we create. While we stipulate in the conditions of the scheme that material must be 'preservation-ready' we often need to assist with the creation of metadata and this can add many months, or even years, to the project.

One of the issues that we are increasingly encountering is the size of the collections. As storage has become more accessible broadly, and it's easier to create digital material, the size of these types of digital collections has of course grown. Community groups have often not undertaken critical appraisal of these collections and want to deposit the collections as a whole. The DRI publishes the collections it preserves under a variety of open-access licenses. Community groups don't always have enough information about the copyright of material they hold so assessing and working through copyright can take time. While we ask that groups have some metadata for the collections they'd like to

[7] Digital Preservation Coalition ‚Bit List', 2022, https://www.dpconline.org/digipres/champion-digital-preservation/bit-list

deposit, it often needs work to get it to a standard that we can accept and this can mean offering more training and support. Supporting groups to overcome these challenges can raise the issues of competing resources at an organizational level as we can find ourselves offering a lot of assistance to those who win the scheme. While we have balanced that in previous years, we are also aware that we can't accept all the organizations who apply to the scheme because we are limited in terms of time, staff and resources. It's important to ensure that the support we offer these groups doesn't impact what we offer our fee-paying members who support the scheme. We are also aware that we need to work to promote the positive opportunities created by the scheme so that fee-paying members also feel invested in how the scheme is progressing and its outcomes.

In Autumn 2021 we developed a training program aimed at community archives and members. Working with the community archives in a more collaborative training setting, and alongside regular members, provided us with important insights about what we need to do to break down barriers. Our Education and Outreach manager captured her takeaways from the sessions, and the thoughts of some of the groups involved in a DRI and DPC blog post 'Breaking down barriers to digital preservation through training'.[8] Including regular members in these training sessions has meant that they get a real sense of the vulnerability and value of the material that is being preserved through the scheme. We are exploring how running a similar type of training program on digital appraisal for community groups might work to make the size of the collections they want to deposit more manageable.

The main obstacle that we have encountered as an organization is that our membership policies are focused on preserving material for organizations we expect to be active well into the future, such as higher education institutions. All of our policies, including our Organizational Manager Agreement are focused on the idea that we steward collections and that where decisions need to be made about the future use of collections, for instance, that decision is made by the depositing organization. Some of the community groups we work with come to us to deposit material because their organization is winding down. Others cannot commit to long-term involvement for various reasons, including a shortage of time and volunteers. While some want to be a core part of our community, others need to be able to deposit and leave.

We began talking in 2022 about how we might restructure the scheme to make it easier for groups to deposit their data. This would mean the introduction of a one-off deposit agreement where we would invite applications from community groups and they could deposit collections outside of the membership framework. We will continue to work with groups who have the material they want to deposit over the medium term but an agreement like this allows us to make organizational plans for how the material will be managed if the community group winds up. There is a balance to be struck here, however. We need to make sure that we create enough space for community groups to be active, own their data, participate and provide input into DRI and our community as a whole, while also allowing them to leave knowing that their data is safely preserved. A second potential route for the preservation of community archive data is through a partnership scheme with current members. This year we are undertaking a pilot scheme to pair community groups with some of our members which will see a community group ingest through a geographically-linked, or disciplinary-linked, member. In this way, the member can become the collection's custodian in the long term, while the community group is recognised as the work's creator.

We are also beginning to develop more partnerships with Community Archiving groups,

[8] Deborah Thorpe, Digital Repository of Ireland, 'Breaking down barriers to digital preservation through training' https://www.dpconline.org/blog/wdpd/wdpd2021-thorpe, https://dri.ie/news/digital-preservation-community-archives/

other archives and funding bodies who support these types of activities as well as looking at how this work can feed into our research project. In September 2023 we are running an Irish Community Archive Symposium with the Irish Community Archive Network (iCAN) that will look at Digital Preservation. iCAN has worked with twenty-eight community archives across the Republic of Ireland. This partnership will allow us to broaden our audience while also raising the profile of Digital Preservation. We are also working with the Radical Archives network this year, which is made up of volunteer and community archivists with digital material. We hope we can work to help support the group in the future. Our work with community groups on the Archiving Reproductive Health project helped us develop a resource for Community Groups 'Guide to Archiving for Community Groups'.[9]

5. Conclusion

The Community Archive Scheme has undoubtedly brought several successes to DRI by diversifying and enriching our datasets and community as a whole, but it has also raised challenges for us as an organization. With a growing network of members and ever-growing data sets, we want to continue to work with these groups in a way that is sustainable for us as an organization but also equitable and supportive for community groups as well. This means recognizing where these organizations do not fit within our regular structures and creating new pathways for these groups so that we can all work together to preserve their digital collections. We also feel that in opening up these new pathways, whether it's through a single depositor scheme or by creating partnerships with our members, new opportunities will arise that we haven't yet considered.

6. References

[1] DRI Membership, DRI website: https://dri.ie/membership

[2] List of DRI Community Archive Scheme winners, DRI website. https://dri.ie/dri-community-archive-scheme

[3] The 'Bit List' of Digitally Endangered Species, DPC website: https://www.dpconline.org/digipres/champion-digital-preservation/bit-list

[4] 'Breaking Down Barriers To Digital Preservation, 3 November 2021, Deborah Thorpe', DPC website https://www.dpconline.org/blog/wdpd/wdpd2021-thorpe

[5] Irish Community Archive Network, National Museum of Ireland website, https://www.museum.ie/en-IE/Learning/Irish-Community-Archive-Network

[6] Digital Repository of Ireland, Archiving Reproductive Health, & Archiving the 8th. (2023) Guide to archiving digital records for volunteer and community groups, Digital Repository of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.k069p160j

[9] Digital Repository of Ireland, Archiving Reproductive Health, & Archiving the 8th. (2023) Guide to archiving digital records for volunteer and community groups, Digital Repository of Ireland [Distributor], Digital Repository of Ireland [Depositing Institution], https://doi.org/10.7486/DRI.k069p160j

ARCHIVER PROJECT: A SUCCESSFUL PUBLIC-PRIVATE COLLABORATIVE PROJECT

Antonio Guillermo Martinez
LIBNOVA SL
Spain
a.guillermo@libnova.com

Maria Fuertes
LIBNOVA SL
Spain
mfuertes@libnova.com

Abstract – The ARCHIVER Project has brought together customers, vendors, and infrastructure providers in an outstanding successful public-private collaborative project, which has also been recognized with the *Award for Collaboration and Cooperation* which celebrates significant collaboration across institutional, professional, sectoral and geographical boundaries at the Digital Preservation Awards 2022.

This paper will review, from the perspective of one of the project winners, the success story of the ARCHIVER Project, highlighting the benefits achieved by leveraging the commercial digital preservation solutions for scientific research data through a pre-commercial procurement process, where end users were able to directly influence the expected functionalities in the platform and how they are expected to operate.

Keywords – Digital Preservation, Research and Development, Collaboration, public-private.

Conference Topics – We're All in this Together.

INTRODUCTION

The ARCHIVER Project (Archiving and Preservation for Research Environments) is the only European Open Science Cloud (EOSC)-related H2020 project focusing on commercial long-term archiving and preservation services for petabyte-scale datasets across multiple research domains and countries [1].

On 29 January 2020, the ARCHIVER project launched its Pre-Commercial Procurement Request for Tenders [2] with the purpose to award several Framework Agreements and work orders for the provision of R&D for hybrid end-to-end archival and preservation services that meet the innovation challenges of European Research communities, in the context of the European Open Science Cloud.

COLLABORATIVE PROJECT

The ARCHIVER project is a clear example of public-private collaboration. Four of Europe's leading research organizations: CERN, EMBL-EBI, PIC/IFAE, and DESY formed a consortium to launch this project in which R&D was performed competitively by commercial providers LIBNOVA and Arkivum [3], through different implementation phases.

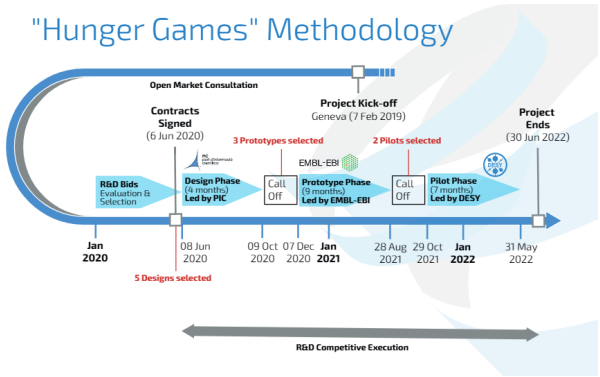
In the case of LIBNOVA, the public-private collaboration of the project was twofold, as in the first phase companies/organizations were invited to combine their skills and resources to form viable consortia to achieve the required results.

In this context and based on this recommendation, LIBNOVA formed a Consortium [4] that was enriched throughout the project with the incorporation of new members with expertise in the specific needs of each phase, forming a multidisciplinary cooperative and collaborative team, combining public sector organizations such as the University of Barcelona and the Spanish National Research Council (CSIC), with private consulting, infrastructure and cybersecurity companies such as Giaretta Associates, Amazon Web Services, Voxility, and Bidaidea.

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

THE R&D METHODOLOGY AND PHASES

In the Pre-Commercial Procurement model, R&D is divided into three phases (design phase, prototype phase, and pilot phase). Post-phase evaluations progressively identify solutions that offer the best value for the money and meet customer needs. Following a "Hunger Games" Methodology [5], where firms were selected or qualified for the next phase, or eliminated. This phased approach allows selected contractors to improve their bids for the next phase, based on lessons learned and feedback from buyers in the previous phase.



The work done at ARCHIVER, which has given rise to the LABDRIVE range at LIBNOVA, changes the approach taken to long-term research data management, both in terms of mindset and technology, i.e. what data researchers keep, how to maintain intellectual control of it, and what data stewards need to do to ensure that value can be derived from it in the long term. The companies selected by ARCHIVER promote environmentally sustainable solutions by providing the means to analyze and reduce the carbon footprint in the digital domain (big data centers).

A key component of sustainability is to ensure that the innovation developed during the project has broad exposure to potential buyers within the European research community and other business sectors. To achieve this, the project has initiated an onboarding process to make the resulting services available to early adopters. Making ARCHIVER services available through the EOSC marketplace will give researchers and contracting organizations the possibility to have sustainable access to these services, being able to test them, evaluate their functionality and purchase them with a clear cost model.

The ARCHIVER effort has resulted in services that can be used immediately by the public research sector in Europe. This will immediately expose novel service offerings, relevant to at least 18 pan-European infrastructures, to the 1.7 million European researchers and 70 million science and technology professionals, public and private sectors combined, who are expected to make use of the European Open Science Cloud (EOSC).

LIBNOVA has demonstrated the outcome of the ARCHIVER R&D activity to a wide group of potential users, both of the services developed and their potential for exploitation by the research community in EOSC [6].

LABDRIVE, THE SOLUTION RESULTING FROM THE ARCHIVER PROJECT

LIBNOVA has been the winner over all three phases of the project (design, protype and pilot), producing the LABDRIVE platform as the project result. LABDRIVE is a **Research Data Management** platform, that supports organizations in their data management endeavors.

During the ARCHIVER project, LABDRIVE has been tested and confirmed to work with High Energy physics, Astrophysics, Life Sciences and other types of large datasets (millions of files and tens of PBs) against 176 combinations of use cases, volume tests, researcher needs and organization requirements, **confirming suitability and scalability of the platform for multiple Research Data Management use cases and needs.**

LABDRIVE is cloud-native, allowing Organizations to leverage the public/private cloud adoption if this is an objective. If not, the platform can also be

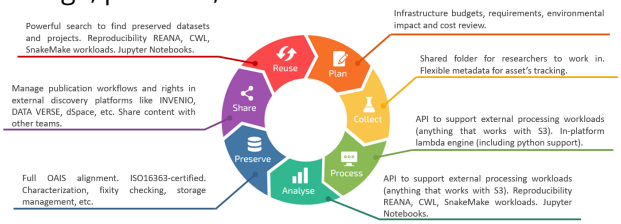
deployed on premises or hybrid cloud/on premises scenarios.

While the LIBNOVA LABDRIVE platform has been re-architected for massive scalability and specific Research Data Management use cases during the Archiver project, LIBNOVA has been the community's trusted partner for digital preservation and data management for several years. Organizations like Stanford University (HILA), Princeton University, Oxford University, The British Library, Pennsylvania State University, Bayer and many other organizations in 17 countries are already LIBNOVA customers.

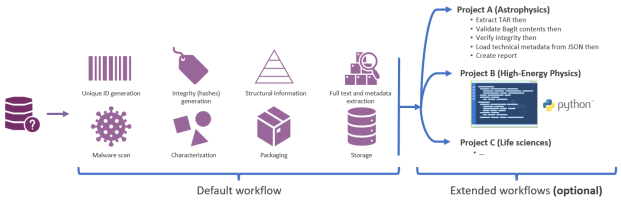
LABDRIVE is a Research Data Management and Preservation platform. It allows organizations to capture the research data they produce, helping them to properly manage, preserve and allow access to it, during the whole data lifecycle.

Design principles.

LABDRIVE provides support over the whole data lifecycle: It allows organizations to capture the research data they produce at the initial stages of the project ("shared folder"), enabling them to properly manage, preserve, reuse and allow access to it:



LABDRIVE works with many research disciplines and content types: It includes a default processing workflow, but it can be extended –using python- to support any other use case. Metadata schemas, data structures, permissions, storage, etc. can also be defined per project, so it can be adapted to multiple scenarios:



LABDRIVE is fully aligned with most relevant and open standards: Fully aligned to the FAIR and TRUST principles [7]. Fully conformant with OAIS [8] and fully aligned with the ISO 16363 [9]. Likewise, ISO 27001, ISO 27017 and ISO 27018-certified. GDPR compliant.

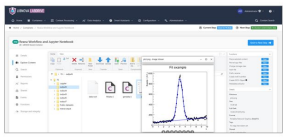
Built on open standards. Aligned with the FAIR and TRUST principles.

Full OAIS/ISO14721 and ISO16363 Alignment. No exit barriers: 100% of the information can be easily migrated.

High security and reliability: ISO27001, ISO27017, ISO27018 certified and externally audited yearly.

GDPR compliant (DPIA for sensitive content).

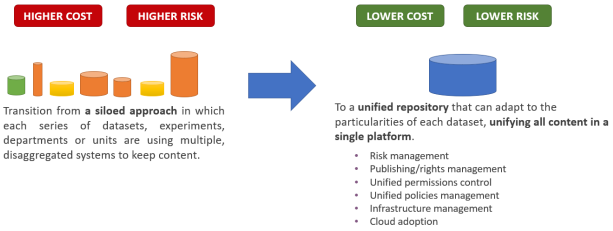
LABDRIVE equally supports power users and simplified use cases: Every action in the platform can be carried out using the easy-to-use web browser interface or the 300-ish Open API methods and 80+ CLI tools available.



Easy to use web interface for users looking for a simplified experience.

Powerful API support, compatible with 80+ CLI tools. Sample Bash scripts, python libraries, etc. available.

As a result, LABDRIVE allows organizations to organize, unify and simplify their research data management strategies, transitioning from a siloed approach to a unified and cohesive platform, obtaining lower risks and lower costs back:



Transition from **a siloed approach** in which each series of datasets, experiments, departments or units are using multiple, disaggregated systems to keep content.

To a **unified repository** that can adapt to the particularities of each dataset, **unifying all content in a single platform**.

- Risk management
- Publishing/rights management
- Unified permissions control
- Unified policies management
- Infrastructure management
- Cloud adoption

LESSONS LEARNED AND BENEFITS FOR PUBLIC-PRIVATE COLLABORATION

Based on the gathered practical experience, a set of lessons learned and best practices can be taken as reference for future PCPs covering aspects such as the procurement process, R&D execution and dissemination of the R&D activities for maximization of results impact by the end of the project [5].

The highlights can be summarized as follows:

- Procurement would benefit if reduced in time and complexity, and focused more on the R&D challenge, as European innovative software SMEs "think" in months rather than years.
- Structured feedback across all parties is found essential, in order to allow full understanding of the challenge.
- The Agile software development methodology can prove to be very effective if

a roadmap for the R&D strategy is produced as a wider frame for expectations, with feature prioritization to avoid possible mismatches in the understanding of the challenge.
- Effort for the tasks of requirement gathering, tender evaluation, assessment and testing of the R&D remains very significant.
- A dissemination plan articulated between the project participants boosts visibility and reach across different communities, sectors and stakeholders.
- PCPs for software services would very much benefit from structured incentives to ramp up the results (for example in the EOSC context), sustaining access to the resulting SaaS and fund trials from researchers in view of purchasing the services if trial deployments are successful.

Overall, the project has demonstrated how the PCP instrument can incite expert SMEs to develop innovative services that can satisfy the needs of Europe's research communities and paves the way to explore more effectively the integration of commercial services into the EOSC marketplace

The work accomplished in ARCHIVER is considered a game-changer for the approach taken to long-term Research Data Management both from a mindset and technological perspective, i.e. what data do researchers retain, how to keep intellectual control of it and what data stewards must do to ensure long-term value can be realized from it.

Thanks to the ARCHIVER Project, the winning companies gain experience from working within public procurement. These are the relevant benefits and tangible results of this collaboration:

- Shorter development life cycle leading to faster time to market, from 5 to 2 years, giving these participating companies an advantage in relation to other competitors.
- Increased customer base portfolio not only in Europe but with contracts signed with universities and other institutions in North America.
- Maximization of the understanding of requirements being able to work with multidisciplinary use cases.
- Pushing the boundaries of what digital preservation is, incorporating innovations that improve the products and empower other organizations to preserve data, consequently creating stronger relationships and incremental business.
- Partnership agreement with hyperscalers (e.g. AWS and Google) strengthening the business perspectives of these European SMEs.
- Increase of services sustainability with a special focus on environmental sustainability.
- Acceleration and de-risking of the ability of these companies to enter a new market with innovative services that address the problem of long-term digital preservation and access to scientific research datasets.

To summarize, the ARCHIVER project has accomplished significant work on technological solutions, its economics and business models, in a holistic manner across scientific domains, public/private sectors and geographies, consistent with the evolving Open Science policies in Europe.

By working directly with the public sector organizations, LIBNOVA and Arkivum were able to receive ongoing input and feedback into their product development to serve the mission of scientific research within Europe, enabling these SMEs to quickly and effectively develop fit-for-purpose products. This resulted in innovative commercialization approaches for the resulting services, improving their degree of FAIRness as an aspect of utmost importance for the ultimate objective of the reuse of research data.

The focus of initiatives such as ARCHIVER is Data, in particular research data, that is set to live for longer than any vendor, system or technology.

Conclusion

The ARCHIVER Project has brought together customers, vendors and infrastructure providers in an outstanding successful public-private collaborative project, which has also been recognized with the Award for Collaboration and Cooperation which celebrates significant collaboration across institutional, professional, sectoral and geographical boundaries at the Digital Preservation Awards 2022 [10]. The award was given to the European ARCHIVER project for what the judges called "important public-private partnership

work that could pave the way for the long-term digital preservation of research data."

The ARCHIVER Project has been a technological breakthrough in the solutions offered by LIBNOVA. In addition to shortening the development times involved in the creation of digital preservation software of the characteristics of LABDRIVE, it allows LIBNOVA to reach market segments that had not been addressed before and to face the design of sustainable digital solutions from a solid position.

REFERENCES

[1] Archiving and Preservation for Research Environments | ARCHIVER Project | Fact Sheet | H2020 | CORDIS | European Commission https://cordis.europa.eu/project/id/824516

[2] ARCHIVER launches its Pre-Commercial Procurement Tender https://cordis.europa.eu/article/id/413444-archiver-launches-its-pre-commercial-procurement-tender

[3] ARCHIVER PROJECT | PILOT PHASE AWARD - THE TWO WINNERS https://archiver-project.eu/pilot-phase-award

[4] ARCHIVER Project | Consortium 3 https://archiver-project.eu/consortium-3

[5] ARCHIVER White Paper. Zenodo. https://doi.org/10.5281/zenodo.7691976

[6] EOSC Marketplace - LIBNOVA LABDRIVE: The Ultimate Research Data Management and Digital Preservation Platform https://marketplace.eosc-portal.eu/services/libnova-labdrive-the-ultimate-research-data-management-and-digital-preservation-platform

[7] LABDRIVE support for FAIRness https://docs.libnova.com/labdrive/concepts/oais-and-iso-16363/labdrive-support-for-fairness

[8] LABDRIVE support for OAIS Conformance https://docs.libnova.com/labdrive/concepts/oais-and-iso-16363/labdrive-support-for-oais-conformance

[9] LABDRIVE - ISO 16363 certification guide https://docs.libnova.com/labdrive/concepts/oais-and-iso-16363/iso-16363-certification-guide

[10] Digital Preservation Awards 2022 – Winners Announced! https://www.dpconline.org/events/digital-preservation-awards/the-winners

DOCUMENTATION GOOD PRACTICE

Bringing Order in Disruptive Times

Jenny Mitcham
Digital Preservation Coalition
UK
Jenny.mitcham@dpconline.org
0000-0003-2884-542X

Abstract – In times of disruption we need to do the less interesting parts of our job better than ever. Documentation falls into this category - a sometimes neglected task that is often sidelined in favor of new and exciting innovations or even just the constant pressure of other routine tasks. It is easy to forget to create documentation or to let existing documents stagnate and become out-of-date. And yet, in the event of a disaster, it may be the very first thing we will turn to, to help to bring order to the chaos. When faced with lockdown as a result of the Covid-19 pandemic in early 2020, there is some evidence that digital preservation practitioners turned to maintenance tasks like documentation when working from home [1]. Digital preservation documentation is undoubtedly important to us in the digital preservation community but where is the good practice guidance that tells us what to document, when, where and how? This paper describes work at the Digital Preservation Coalition to gather together community experiences to create a new good practice guide on digital preservation documentation.

Keywords – Documentation, Good practice, Guidance, Collaboration

Conference Topics – WE'RE ALL IN THIS TOGETHER; FROM THEORY TO PRACTICE

I. INTRODUCTION

In February of 2023 the Digital Preservation Coalition (DPC) began a small project to create a good practice guide to digital preservation documentation. This was a theme that had been flagged up by DPC members more than once as a topic of interest and we were keen to publish a resource containing helpful advice, both for our members and for the wider community. The topic had been raised most often in the context of our Rapid Assessment Model (DPC RAM), a maturity model for digital preservation [2]. In the frequent conversations we have with our Members around the model, questions about documentation often emerged.

The Rapid Assessment Model encapsulates digital preservation good practice and includes examples of activities that should be in place in order to move up to a higher level of digital preservation maturity. For example, at the 'Basic' level of the Policy and Strategy section of the model it is mentioned that "some procedures for managing, and providing access to, digital content are in place and may be documented". At the 'Managed' level it is suggested that "a suite of documented processes and procedures for managing, and providing access to, content within the digital archive exists". Other mentions of documentation appear throughout the model.

The theme of documentation also runs through the NDSA Levels of Digital Preservation [3], with work on the Levels Reboot even going so far as considering adding a new row to the Levels with a focus entirely on documentation. Documentation is mentioned directly six times within the Levels matrix, making its importance quite clear.

In a call to action published on the DPC blog, Amy Rudersdorf of AVP highlighted the importance of documentation as part of any digital preservation program and provided a persuasive list of reasons why we should all focus more time on it [4].

These examples all help to highlight the centrality of documentation to recognized digital preservation good practice, but none of these sources describe *how* we should do it.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

The OSSArcFlow Project which ran from 2017 to 2020 has produced some valuable outputs of relevance to this question. It provides a methodology and a range of examples relating to the documentation of digital preservation workflows using open-source tools. Of particular interest is their Guide to Documenting Born-Digital Archival Workflows [5].

Further examples of digital preservation workflows can be found on the Community Owned Workflows (COW) wiki [6]. There are some good examples here of documented workflows and diagrams which may act as inspiration for those who are looking to create their own documentation.

Outside of the digital preservation community there are further resources that can be accessed to learn more about documentation. Write the Docs describes itself as a "global community of people who care about documentation" [7]. Though much of this resource is focused on documenting code, there are certainly some useful tips to be found that are more broadly relevant to documenting digital preservation processes and procedures.

Good documentation clearly *is* good practice for the digital preservation community but how should we go about this task? It is clear that some useful resources already exist, but it was recognized that a guide to provide advice to practitioners on how to approach their digital preservation documentation challenges would be helpful.

II. Methodology

Collaboration is built into the workings of the DPC, and it seemed an obvious step to bring together a group of practitioners to share thoughts and experiences on digital preservation documentation and brainstorm some of the key questions which would be addressed within the good practice guide.

Volunteers were sought from the DPC Membership to come together in a series of focus groups to discuss the topic of documentation [8]. There was considerable interest in this call and a wide range of organizations expressed a desire to be involved. The focus group meetings took place in February and March of 2023. To accommodate different time zones, two separate meetings were arranged. This led to smaller groups and helped facilitate more inclusive discussions and open sharing of ideas. With the help of sticky notes on a Google Jamboard, and question prompts for discussion, participants were invited to share their thoughts on topics relating to documentation, in particular looking at the five W's (and one H) [9] of documentation:

- **Why?** Why do we document and what are the risks if we don't?
- **What?** What should we be documenting?
- **Who?** Who are we documenting for?
- **Where?** Where should we store our documentation?
- **When?** When should we document, when should we revise and update it and at what point should we preserve it?
- **How?** How should we document and how should we maintain it?

III. Scope

The first task of the focus group meetings was to discuss and agree the scope of the work. Documentation is a big topic, so keeping the scope tight and focused was important in ensuring the task of creating a good practice guide was manageable. It was agreed that the documentation in scope was as follows:

Documentation that is important for the day-to-day operations of digital preservation activities within an organization, for example recording how digital preservation tasks and procedures are carried out or how tools and systems are integrated and configured.

Elements of digital preservation documentation that were considered out of scope were:

- Digital preservation policy or strategy documents - this is a very specific type of documentation, and guidance on this is already well covered (see for example the recently revised Digital Preservation Policy Toolkit [10]).
- Documentation relating to high level planning and reporting – this guide was focused on documentation that describes processes and workflows rather than that which describes and informs future plans.

- Documentation that describes individual datasets to enable them to be understood and re-used - though this subset of documentation is clearly very important, it has quite a different emphasis and purpose to documentation specifically about digital preservation operations.

IV. Discussion

Discussions within the focus groups were lively and interesting and participants had no shortage of ideas and experiences to share. Documentation is a topic that is of relevance and interest to everyone, and it was interesting to learn about different ideas and approaches to tackling this task across different organizations. The question prompts and discussion not only elicited sharing of current practice but also encouraged some participants to consider changes to their own practices as a result of learnings from the sessions. It was encouraging to see positive outcomes such as this even prior to the guide being written.

V. Good Practice Guide

The focus groups provided a wealth of material which could be condensed into a series of helpful sections of the guide. Focus group participants were also able to provide comment and feedback on the draft text for the guide as it was developed and were encouraged to supply examples and written case studies to help to illustrate the advice given.

The main sections of the guide are described below:

- Why documentation is important – this section describes the benefits of documentation (along with the risks if documentation doesn't exist).
- Audiences for documentation – a summary of the internal and external audiences who documentation may be intended for.
- What makes good documentation and what makes bad documentation – this section takes the form of a table summarising some of the key characteristics of good documentation and bad documentation.
- Tips for creating documentation – this section includes information about methods, tools, templates, diagrams and testing.
- How to maintain documentation and manage versions – this section covers the challenges of keeping documentation up to date over time and how version control should be managed.
- Preserving documentation – this section of the guide briefly describes why documentation may need to be preserved for the long term and some of the things that should be considered.
- Case studies – members of the focus groups have provided case studies about their own documentation practices. A range of types of organization were selected, with different tools, platforms and practices represented.
- Examples – some organizations make elements of their documentation publicly available online. The guide shares links to helpful examples which can be used for inspiration.
- Further reading – useful links and references are shared to other resources.

Digital Preservation Documentation: a guide [11] will be publicly launched at iPRES 2023 and freely available for all to consult.

VI. Conclusion

The opposite of disruption is calmness, tidiness, and order. Whether our work in digital preservation is disrupted or not, the presence of well-crafted documentation should provide a level of reassurance in our processes and procedures both now and in the future. It is the author's hope that the guide, released as a result of this collaborative work, will provide helpful advice to the digital preservation community on creating, managing and preserving digital preservation documentation. Good documentation is an essential element of digital preservation good practice and one which should not be put off until tomorrow.

1. REFERENCES

[1] Don't Let All That Work Go to Waste: Documentation Strategies for Success, Nathan Tallman and Carly Dearborn, 2020, doi:10.26207/f4ts-rm53

[2] Rapid Assessment Model, Digital Preservation Coalition, Version 2.0, 2021. http://doi.org/10.7207/dpcram21-02

[3] Levels of Digital Preservation Matrix V2.0, Levels of Preservation Revisions Working Group, October 2019. https://osf.io/2mkwx/

[4] DOCUMENT THIS. And this. And this, too., Amy Rudersdorf, 2019. https://www.dpconline.org/blog/wdpd/document-this

[5] OSSArcFlow Guide to Documenting Born-Digital Archival Workflows, Alexandra Chassanoff and Colin Post, 2020, https://educopia.org/ossarcflow-guide/

[6] Community Owned Workflows, https://coptr.digipres.org/index.php/Workflow:Community_Owned_Workflows

[7] Write the Docs, https://www.writethedocs.org/

[8] Digital preservation documentation - join our focus group! Digital Preservation Coalition, 2023. https://www.dpconline.org/news/dp-documentation-fg

[9] Five Ws, Wikipedia, 2023. https://en.wikipedia.org/wiki/Five_Ws

[10] Digital Preservation Policy Toolkit, Digital Preservation Coalition, 2023. https://www.dpconline.org/digipres/implement-digipres/policy-toolkit

[11] Digital Preservation Documentation: a guide, Digital Preservation Coalition, 2023 - http://doi.org/10.7207/documentation-23

RESCUING LEGACY DIGITAL COLLECTIONS

Lessons Learned from Migrating Historical Bespoke Digital Collections

Kayla Maloney
The University of Sydney
Australia
kayla.maloney@sydney.edu.au
0000-0001-6247-3944

Katrina McAlpine
The University of Sydney
Australia
katrina.mcalpine@sydney.edu.au
0000-0002-2305-3661

Jennifer Stanton
The University of Sydney
Australia
jennifer.stanton@sydney.edu.au
0000-0002-6285-1340

Abstract – The University of Sydney Library hosts many historically significant digital collections. In 2021 and 2022, the Library undertook a project to ensure the accessibility of these collections, migrating them from ageing web servers to our current repository systems. This paper outlines the challenges involved in managing bespoke legacy collections at an institution in the early stages of building digital preservation capacity. We discuss the approaches taken to make use of existing systems, capabilities, and resourcing to rescue collections and prepare for future preservation actions.

Keywords – Digital humanities, Legacy digital content, Data curation, Sustainability, Digital preservation

Conference Topics – We're All in this Together; Sustainability: Real and Imagined.

I. INTRODUCTION

The University of Sydney Library was an early adopter in creating and supporting digital cultural collections. The Library has been hosting online digital collections since 1996, and by 2021 had on the order of 85 different collections being hosted on 15 servers.

Content across the collections varied widely. The collections included historical photographs, digitized manuscripts and images from Rare Books and Special Collections, transcriptions of handwritten content, an archive of archeological grey literature, artworks produced by staff and students from the University's Sydney College of the Arts, and an archive of audio files of Australian adolescents' speech from the 1960s, to name a few. The collections comprise historically relevant content, particularly in an Australian context, and document early digital humanities projects and experiments in using technology and online display in novel ways.

This paper discusses a project to migrate these collections to more modern systems, keeping this historic content accessible and usable for the future, without having a mature digital preservation program in place. We discuss some of the challenges encountered working with legacy collections and infrastructure. We hope that our project can provide insights for people working with non-standardized, bespoke content where there may not be an obvious "right" way forward.

II. BACKGROUND

Despite the experimental nature of several of the collections, little intervention from Library staff was required to keep them online and available over the decades. Consequently, a lot of the institutional knowledge around the collections was gone by the time the Library started this migration project

Library staff have been exploring issues around these legacy collections and how they should be managed since 2017. However, getting a comprehensive picture of the entirety of our content was not straightforward. To save the cost of setting up additional servers, new collections were often added to existing servers, resulting in a complex web of links and sometimes orphaned pages. Some of the servers were originally physical, virtualized years later, and finally, years later again, were moved to the

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

cloud. They were beyond their end-of-life and no longer fit for purpose.

To properly tackle this situation, we needed someone with the appropriate technical skills to dedicate a large amount of time to investigate the collections and determine appropriate solutions for different cases. However, Library IT staff were in high-demand and there were few staff with the skillset needed to navigate the ageing servers. Over the years, at least four different people started to investigate and audit the content on separate occasions, only to be pulled away when urgent tasks elsewhere required attention.

During this time, the Library began to invest in a digital preservation program. Staff undertook training and development activities, including iPres conference attendance, the Digital Preservation Coalition's 'Novice to Know How' course, completing digital preservation maturity modelling and implementing some digital preservation workflows for digital collections. Overall, however, digital preservation at the Library was still in its infancy and a digital preservation framework or system had not been implemented.

In April of 2021, rising institutional cyber security concerns led to a deadline for upgrading or shutting down the legacy collection servers. This was no longer a task that could be put on the backburner until we had the time to do it "properly".

Staff from the Digital Collections team, Library IT and the Sydney University Press compiled a comprehensive list of collections from the legacy servers, based on the earlier audit. The team looked to projects at other institutions on managing and preserving bespoke digital humanities collections to develop approaches for rescuing and migrating the content in our collections [1]. Each collection was assessed for whether it should be kept, and where and how it should be migrated. Tasks were assigned to the appropriate team, and everyone got to work.

III. Challenges

A. Have we found everything?

Gaining a comprehensive understanding of all the collections on our servers had been a major roadblock to getting started on this project for years, and the worry that we might be missing something was with us throughout the entirety of the project.

To ensure that we had a copy of all content, the final step in decommissioning each server was to archive all content and configuration files and put the archive on the University's Amazon Glacier storage. Concerns emerged at one point that two of our more unstable servers could fail before being properly decommissioned. Due to staff availability, we were unable to undertake priority archiving of these servers according to our established process. As a stop-gap measure, team members attempted to use the MacOS application SiteSucker to get a local emergency backup copy of these servers [2]. This was successful for one of the servers, but SiteSucker struggled to capture the entirety of our most complicated server, and we were left with an incomplete emergency backup. Fortunately, both servers remained functional until they were able to be properly archived and decommissioned.

These backups mitigated the risk of data loss, however, they did not solve the problem of knowing what content we needed to migrate, and understanding how that content displayed and functioned in its original context.

Where SiteSucker worked, it provided us with the additional benefit of easily accessed working copies of our content and insights into where we had content that we had not yet identified. We also manually combed through the sites and tried web searches to turn up orphaned pages still hosted and accessible, but no longer linked to from the main pages of the sites. Some orphaned pages were only discovered through serendipity, for instance, a team member finding a reference to a collection in historical documentation, or an inquiry from a member of the public. These finds helped us move forward, but also highlighted the likelihood that we were missing content from our migration plan.

For websites that hosted large numbers of files available for download, such as PDFs, we used the browser extension Simple Mass Downloader to obtain local working copies of the files for migration [3]. This was also helpful for cross-checking with existing and newly created collections metadata, to highlight gaps where we might be missing files or where we needed to create metadata.

Our intention was to migrate content with no downtime, so that the new location would be available prior to removing the old. We eventually reached a point in our checks where we felt the risk

of downtime due to missing a collection was acceptably low, and our backups gave us confidence that we would be able to reinstate any content that we missed.

B. Understanding our content

Documentation was uneven across the legacy collections. For some, it was difficult to determine important details such as the copyright owner, agreements that had been made around the collection, who had been involved, or sometimes even why we had it in the first place. This information can be critical in making decisions about what preservation actions can or should be taken for a collection. Statistics around usage and engagement with the different collections would also have been valuable for this decision-making, however, issues with the setup of the servers and the influence of bots meant that we were unable to get trustworthy information.

Interestingly, the fact that many of these collections had continued to remain accessible with minimal intervention over long time periods was a contributing factor to the loss of institutional knowledge. Most of the bespoke collections were built using HTML and we did not need to grapple with the complex issue of preserving custom software. Without problems occurring, no one needed to check in on the collections and staff who had been involved in collection creation left the institution without passing on historical knowledge. For most of the collections, particularly those where the Library was involved in their creation, we were able to turn up the information needed. This took the form of finding historic documentation, relying on institutional memory from some long-term staff, or tracking down contact details from involved parties. In a few cases, the information that we found allowed us to determine that we no longer would make a collection available, for instance, where an agreement had lapsed, or if the purpose that it was made available for was no longer relevant. In some cases where documentation was lacking, we had to decide whether the Library was the best organization to make content available. Other institutions have subsequently digitized some of the same materials at a higher quality. When better versions were openly available elsewhere, we generally opted not to migrate our version.

In all cases, we tried to ensure that the information we turned up and any decisions we made were well documented. Project decisions were recorded in project documentation. Where investigations were required, outcomes from the investigation were detailed in Word documents and stored alongside collection files in our dark archive location. Agreements regarding collections were saved to the University's recordkeeping system and the record numbers were added to administrator metadata for the collection in our repository systems to ensure connection between the information across the systems. A brief statement about the migration was added to items' provenance metadata fields in their new location, visible only to system administrators. We also considered how best to include information for others to use and understand the collections. "About" pages were created detailing the projects that many of the collections belonged to, outlining the history of the projects, funding, references to agreements around collection content and an acknowledgement of the people involved. We also included a link to versions of the sites archived in the Internet Archive's Wayback Machine to allow people to see the original context of the collections. These pages are hosted on our current Digital Collections site.

C. Non-standard structures and scale

The Library no longer hosts servers for individual digital collections to have their own bespoke pages. Instead, we have moved towards having more standardized systems and processes, including the Library's Digital Collections repository [4] (Recollect [5]) and the Sydney eScholarship repository [6] (DSpace [7]) for University research outputs. As repositories, these systems have different affordances to websites. It was not always straightforward to determine how the bespoke, and frequently unusually formatted, website-based content should best be migrated and displayed in a repository system.

The John Anderson Archive provides an example of one of our approaches to unusually formatted content. The Archive presents significant works and papers of John Anderson (Challis Professor of Philosophy at the University from 1927 until 1958) [8]. Among these works are handwritten lecture notes. The original form of the Archive presented transcriptions of the notes as HTML text on the

website. Each transcribed page included a link to an image of the original handwritten text. The handwritten notes also included asides, often indicated by text in square brackets. The asides were included in the transcriptions as hyperlinked notes that opened in a separate pop-up window. Significant reformatting was needed to be able to include this content in our Digital Collections repository. The transcribed text was copied and pasted into a Word document, preserving the page numbering of the original text. The hyperlinked notes were included as footnotes. Each document was saved as a PDF and uploaded to the repository. The JPEG images of the original handwritten notes were combined and saved as a PDF and uploaded to the repository as a separate item to the transcriptions. In this way, we were able to preserve the content of the original archive, although not the rather experimental functionality of the linked notes and images. Due to the manual nature of this work, significant resourcing was required. We benefitted from the availability of additional staff, who normally work in client-facing roles, during lockdowns and periods of reduced services during the COVID-19 pandemic.

Other collections were large enough that a manual approach was not feasible. A collection of archeological reports [9] and another of photographs of artworks produced at the University's Sydney College of the Arts [10] each contained well over 1,000 items. No reformatting of the content was needed, however, collection item metadata needed to be combined, mapped, and transformed for ingest to our Digital Collections repository. The artwork metadata was originally stored in a relational database, where many images, each with their own metadata, could belong to a single artwork. We needed to transform this to a flat tabular structure. To do this, we used the pandas Python library for the data wrangling and Jupyter notebooks to allow us to document our code in a more readable fashion for future reuse. We also took the opportunity to involve team members with no coding experience to enable knowledge-sharing and the development of new skills across our team.

D. Digital preservation maturity

The Library was, and at the time of writing still is, in the early stages of implementing a digital preservation program. Ideally, we would have undertaken this migration project with a more mature digital preservation program and an appropriate digital preservation system in place, however this was not an option. Throughout this project we were able to apply some digital preservation practices such as using tools like TeraCopy to transfer files, ensuring there were back up files created and stored and that the project was well documented. However, we were, and are, aware that there were many processes we could not complete due to lack of time and an established preservation framework. This was challenging, as we knew throughout the project that there were digital preservation good practices we were not following, and that there would be extensive future work to undertake to enable us to move our content into a digital preservation system.

IV. Lessons Learned

Our main lesson from this work is that we cannot let the desire for a perfect solution prevent us from getting started. We do not want to go in and start doing work without considering issues and having a plan, but if getting that plan completely "right" means important work never gets started, we need a different approach. Not all issues can or should be solved upfront, and we can work through problems as they come. This may lead to stress when something unexpected crops up, or we realize that we have overlooked something; not everything will be done in the ideal way. Even with these bumps along the way, it is a far better outcome than never getting started and losing everything.

Documentation is critical for being able to appropriately manage and preserve content, but historical practices have not always given us the information that we need. This includes information about copyright holders, agreements and reuse conditions, project stakeholders, and collection outcomes and impact. Tracking this information down can take a lot of resourcing. Where needed, taking a risk-management approach can help us to make acceptable decisions. Whatever happens, it is essential to set ourselves up better for the future by documenting this important information and what we have done using the tools available to us, including recordkeeping systems, collection metadata, project histories and project documentation.

We also learned that we should consider whether content, functionality or both need to be retained when migrating to new systems. Our systems did not always allow us to preserve the functionality of the content we migrated, however, this web-based content had been archived by the Wayback Machine, allowing us to link to earlier versions to provide users with the initial context for the collection.

Collections may be hosted in one place, but over time, they will be harvested and linked to elsewhere. Any time collections move, issues will appear in the network of places they now exist in. Permalinks can help to mitigate this issue, but they will not entirely solve it. Issues can be chased down over time as they are noticed, and this should be seen as something to be aware of, but not something that we can fully plan for from the beginning of a project.

Finally, a project like this will require a large range of skills to complete. Wherever possible, we tried to prioritize and make the time to share knowledge and skills. This will mean that some tasks take longer than if the staff member with the most knowledge completes them fully. Particularly in areas where only one staff member has a skill, the growth in team capacity is well worth this extra time.

V. Future Considerations

This paper has outlined a project to rescue legacy collections from being lost entirely. The current systems that they have been migrated to are repository systems that enable access but are not preservation systems. The University of Sydney is increasingly interested in digital preservation, and there is likely to be future institutional support for growing our digital preservation capacity. The actions taken to standardize collections in this project will assist us in future preservation activities and working with future systems.

Digital humanities projects and bespoke digital collections similar to those addressed by this project are still being created. Migrating the collections has given the Library and the University further insights into what needs to be considered for managing these projects and outputs in the future. Do we need to be creating service level agreements for ongoing support of collections? What information, agreements and documentation do we need to have to ensure that we can manage and preserve a collection throughout its life? What constitutes end-of-life for a collection or project, and what should happen next? These are some of the questions that we are grappling with as we plan our future digital preservation program.

VI. Acknowledgements

The authors would like to acknowledge everyone involved in this collection migration project, including Sarah Graham, Susan Murray, Susan Brazel, Phil Jones, Marthe Follestad, Piyachat Ratana, Ryan Stoker, Arin Bryant-Munoz, Dora Zhang, Wayne Zhang, Keerat Judge, and Rengen Parlane. Thanks to Nicholas Keyzer and Anthony Green from Schaeffer Library for providing us with images and metadata for the Sydney College of the Arts Archive. We would also particularly like to thank Jim Nicholls and Kim Williams, who both played major roles in the project and provided us with information and feedback on this paper.

1. REFERENCES

[1] J. Smithies, C. Westling, A.-M. Sichani, P. Mellen, and A. Ciula, "Managing 100 Digital Humanities Projects: Digital Scholarship & Archiving in King's Digital Lab," *Digit. Humanit. Q.*, vol. 13, no. 1, 2019, [Online]. Available: http://www.digitalhumanities.org/dhq/vol/13/1/000411/000411.html

[2] "SiteSucker for macOS." https://ricks-apps.com/osx/sitesucker/index.html (accessed Mar. 03, 2023).

[3] "Simple mass downloader - Chrome browser extension." https://chrome.google.com/webstore/detail/simple-mass-downloader/abdkkegmcbiomijcbdaodaflgehfffed (accessed Mar. 03, 2023).

[4] "Digital Collections | University of Sydney Library." https://digital.library.sydney.edu.au/ (accessed Mar. 03, 2023).

[5] "Recollect - Collection Management and Community Engagement Software." https://www.recollectcms.com/ (accessed Mar. 03, 2023).

[6] "Sydney eScholarship Repository." https://ses.library.usyd.edu.au/ (accessed Mar. 03, 2023).

[7] "DSpace." https://dspace.lyrasis.org/ (accessed Mar. 03, 2023).

[8] "John Anderson Archive." https://digital.library.sydney.edu.au/nodes/view/6932 (accessed Mar. 03, 2023).

[9] "NSW Archaeology Online." https://digital.library.sydney.edu.au/nodes/view/6929 (accessed Mar. 03, 2023).

[10] "Sydney College of the Arts Archive." https://digital.library.sydney.edu.au/nodes/view/6927 (accessed Mar. 03, 2023).

QUALITY PRESERVATION

Emerging Quality Assurance Practices in the Library of Congress Web Archives

Meghan Lyon
Library of Congress
USA
mlyon@loc.gov

Grace Bicho
Library of Congress
USA
grth@loc.gov

Abstract – Building sustainable quality assurance practices is a challenge for today's preservationists, who want to be sure that content preserved in web archives is not only the correct content, but in working order. This often means that archived web content should be replayed via Wayback rendering software in good fidelity when compared to the original website. The exponentially growing scale of web archives necessitates a multipronged approach to identify what is (and is not) being preserved, and where improvements can be made. This paper will explore actions that can take place iteratively throughout the web archiving life cycle, as part of a larger system of review where multiple individuals can contribute, including non-technical Library staff and subject matter experts. The processes described are part of a novel workflow in the Library of Congress Web Archiving Program.

Keywords – Web Archives, Quality Assurance, Workflows, Human-centered digital practitioners

Conference Topics – From Theory to Practice, Sustainability, We're All in this Together

I. INTRODUCTION

The Library of Congress Web Archiving Program manages an ever-growing archive of over 3.5 Petabytes (PB) of content archived from the web since 2000. The archive comprises over 180 event and thematic collections, nearly 31,000 cataloged web archives, and approximately 15,000 seed URLs ("websites") actively crawling at any given time. The Library's technical Web Archiving Team (WAT) is responsible for managing the program from start to finish, which includes leading the assessment of archive quality, even though the WAT does not select content for the archive.

Assessing the quality of web archives is a notoriously difficult endeavor for the web archiving community, given the sheer chaos of file formats present in the archive, the quickly increasing scale, and persistent replay issues with the current suite of access tools, which will always lag behind new technologies used to build the live web. However, it is seen as due diligence by the WAT to confirm capture of selected content for the Library's collection. WAT also approaches quality assessment as an act of sustainability, within the feedback loop of the Library's ongoing captures, in order to scope capture to *only* content that has been selected for the collection, according to the Library of Congress Collection Policy Statements [1]. Finally, performing quality assessment allows the WAT to provide a reasonable expectation of the usability of the archive for those building and using the collection [2].

This paper presents a detailed explanation of the Library of Congress Web Archiving Team's practical approach to quality assessment of the web archive, including computer-mediated methods, according to Dr. Brenda Reyes Ayala's theoretical framework for performing quality assessment on archived web content [3].

II. THEORETICAL FRAMEWORK

The "human-centered grounded theory" [3] is the first of its kind to provide a theoretical framework for increasing web archivists' confidence in quality assurance (QA) methods in the face of the enormous scale of managing web archives. The grounded theory includes three dimensions used to assess quality of the web archive: Archivability, Relevance, and Correspondence.

A. Theoretical Definitions

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

1) *Archivability:* "the degree to which the intrinsic properties of a website make it easier or more difficult to archive."
2) *Relevance:* "the pertinence of the contents of an archived website to the original website. Reference [3] defines two measures of relevance: topic relevance and size relevance."
3) *Correspondence:* "the degree of similarity, or resemblance, between the original website and the archived website." Reference [3] defines three measures of correspondence: *visual correspondence, interactional correspondence, and completeness.*

III. Archivability

Archivability is the most difficult dimension to assess completely as website-building frameworks are constantly changing, and web archiving technology is slow to adapt. The WAT works with its vendor, who performs the data capture component (known as web "harvesting" or "crawling") of the web archiving life cycle [4], to begin assessing archivability. The WAT also takes on the responsibility of communicating archivability to nominators—non-technical Library staff responsible for selecting content for the archive—in order to manage expectations of what is possible to archive.

A. Vendor Collaboration

The Library's crawl vendor works continuously to improve the captures of selected content and to determine which web development technologies make crawling difficult. Before a harvest begins, the vendor first uses a technology, such as Wappalyzer [5], to scan a website for frameworks, programming languages, web servers, and anything else that may impede capture. Based on the results, the vendor can decide which crawling technology is best suited to harvest each website. Once a crawl finishes, the WAT can provide feedback about how well the technology worked, and can suggest movement among various crawl technologies. This collaborative feedback loop is critical in identifying challenges with archivability.

B. Known Challenges

Over time, working with the vendor and assessing crawls, the WAT has built up a list of common challenges with certain platforms or websites. In order to manage expectations of crawling and archive replay for nominators, the team provides a table of guidelines, on an internal Wiki, called "Web Archiving Known Challenges." Nominators are then able to consult the list at any time, particularly during initial content selection or while assessing crawl quality of their selected content.

IV. Relevance

According to [3], the core category of relevance is split into two dimensions: topic and size relevance. Topic relevance measures the closeness of a web archive to the original, live website or part of a website. This curatorial measurement is largely outside of the scope of practice for the WAT. The second dimension of size relevance, or how closely a web archive's size correlates to the live website, is within scope for WAT, the technical team tasked with assessing quality of incoming web harvests.

Since it is difficult to determine the size of any given website, it is also difficult to determine whether the size of the archived version matches the live website. Some web archiving programs run test crawls to determine archivability and accuracy of crawl instructions, and are able to determine approximate website size at that point. However, the Library only crawls at ongoing, regular intervals, providing the ability to compare the size of archived versions over time, as well as identify websites that appear unreasonably small or unreasonably large, given the number and types of resources it takes to make up a website.

Using reports generated by the crawler software and crawl vendor, the WAT devised a method for assessing the relative size of each seed (or website URL at which the crawl is set to begin harvesting). The reports utilized are: the Heritrix crawler standard seeds-report.txt report [6], including the response codes and HTTP status of each seed at the time of harvest, and a bespoke report of the number of hops traversed (or depth) and number of raw bytes collected (or bytes) per seed by the end of each crawl.

The above data points are collated by the WAT into a spreadsheet and are matched with collection data from the program's curatorial database per seed URL. From there, the WAT can easily sort by the response codes, depth, and bytes, or by a particular collection or crawl frequency.

Various sorting highlights initially the websites with extremely low bytes and depth that had obvious crawl issues. From there, the WAT staff performing QA can triage the investigation of seeds with low- to mid-range bytes and depth as an indication of difficulty crawling some or all parts of the seed. Resolutions of these investigations can look like switching the crawl technology for a particular seed, updating crawl instructions (or "scopes") for the web crawler, or removing the seed from crawl altogether.

In this way, the WAT leans into the iterative flow of the Library's unique crawling ecosystem, using relative size of the seeds in a crawl and over time to highlight acute seed issues.

V. CORRESPONDENCE

The WAT is responsible for overseeing the capture of approximately 15,000 seeds at any given time. Regarding the assessment of quality for those seeds, archivability and size relevance help immensely to highlight seed issue needles in the archive haystack. To look deeper into the quality of each site at scale, subject expertise and the measures of correspondence come into play, a process which the WAT calls "capture assessment."

A. Capture Assessment: Data Collection

For the Library, all three correspondence categories: visual correspondence, interactional correspondence, and completeness, rely on the nominator's knowledge of the live website for comparison. To gather actionable information about quality from nominators and other staff supporting review of the content–referred to as "reviewers" in the context of performing capture assessment–WAT has translated the three categories into a rubric to be measured. For each category, a numeric range is instituted from 1 (worst) to 5 (perfect), which the reviewer can use to ascribe a numeric value for that category for a single capture of a seed.

The visual correspondence score can range from appearing "unrecognizable" (1) to appearing "perfect" (5). The WAT's prompt elaborates, "similarity in appearance between the original website and the archived website" [3] by asking reviewers: *If you were to look at the archived page and the page on the live web side by side, how similar would they look?*

Similarly, the interactional correspondence category includes the definition, "the degree to which a user's interaction with the archived site is similar to that of the original" [3], alongside a series of questions meant to flesh out the concept, such as: *Do the navigation buttons function? Is there an endless scrolling feature or interactive visualization?,* and *Does it work in the archive?* The interactional correspondence score can range from inability "to interact with any features of the archived website" (1) to ability "to interact with all features of the archived website" (5).

Completeness, "the degree to which an archived website contains all of the components of the original", asks reviewers to *get a holistic sense of the archive. What overall patterns emerge as you navigate around the archived site?* We ask reviewers to rank the whole capture to say "no content missing" (5), "some content missing" (4), "half content missing" (3), "most content missing" (2), and "all content missing." (1)

If the rating of any category is any less than 5, the WAT provides a checklist of common issues that communicate the issue they are seeing with that capture, including a free-text "other" box for any unlisted issues. Some of the common issues in the checklist include: *Missing images*, *Missing documents*, *Missing style*, *Paywall or login impedes use*, *Page elements disappear*, and *Issues with interactive content*.

An introduction to the work of Reference [3], a rubric for correspondence scores, and a Specific Issue checklist is presented to the reviewer within a Confluence form. When the form is submitted, WAT gets an email with the results and can act on identified issues. However, in order to streamline review of capture assessments, WAT exports the form results at regular intervals, integrating work reviewing the capture assessments with bi-weekly work-planning sessions within the team's Scrum workflow [7].

B. Capture Assessment: Action steps

Individual tickets are created, per capture assessment form response, in a workflow organizer (Jira) and assigned at random to WAT staff. Before importing into Jira, the form response data undergoes a transformation via Python script. This step fulfills the dual purpose of: 1) formatting the form responses into an order suitable for bulk-

import to Jira tickets and 2) averages the 1-5 correspondence ratings. The average of the three correspondence ratings dictates the priority level of the Jira ticket:

1) *Blocker:* a score of 1 in any category
2) *Critical:* average correspondence score less than or equal to 2
3) *High:* average correspondence score greater than 2 and less than or equal to 3.5
4) *Medium:* average correspondence score greater than 3.5 and less than 5
5) *Minor:* average correspondence score of 5, exactly, indicating a perfect capture

Prioritization of quality assurance is critical in web archives, which have endless opportunities for improvement, but real human limits. Assigning Blocker to a given capture assessment ticket indicates to the WAT that a crawled seed requires attention immediately. A Medium score, on the other hand, is indicative of something wrong, which can often be righted with a small adjustment by WAT, such as updating the crawl instructions.

C. *Early Results*

Six months into the effort to put theory into practice, WAT is beginning to see preliminary results. Over 193 captures of seed URLs have been assessed by 15 unique reviewers across 13 collections (some collections had multiple reviewers and some unique reviewers assessed captures from more than one collection). An average correspondence score of 3.86 has emerged. By priority, roughly 30% of tickets land in the Blocker, Critical or High priorities with the remaining 70% at the Medium and Minor levels. During February 2023, the WAT averaged 7.5 days to complete processing of new capture assessments.

The majority of assessments (54%) were performed on content collected as part of a multi-disciplinary, cross-divisional collecting effort geared toward collecting publications via web archiving. This collection is unique to the Web Archiving Program in that it has acquisitions staff assigned to the collection who act as liaisons between staff with recommending authorities and the WAT. In keeping with the collection's focus, the most widespread specific issue discovered in this collection is *Missing documents* (38% of all reported issues for the collection), followed by *Missing content* (other) and *Missing links* (11% each).

Of the 310 specific issues reported across the assessed collections, the highest counts of specific issues checked were *Missing images* (21%), *Missing documents* (19%), and *Missing style* (13%), which is a common formatting error where CSS is either not captured or improperly rendered.

It is helpful for reviewers to indicate when they see something "missing" that they expect to be present in the archived capture. Reviewers with language and subject expertise highlight areas of the site most critical to collect. When these specifics are pointed out, WAT can investigate further to verify whether something is truly missing from the archive versus un-navigable from a given starting point, thereby ensuring capture of content selected for the Library's collection.

Investigation often begins by consulting the live site for the URL in question, or a representative URL of the larger issue, i.e., an image URL if *Missing images* was checked. With a URL in hand, WAT can pinpoint examination of the resource via Wayback replay or the archive indexes to better understand whether the URL is truly absent in the archive. WAT can then compare the document URL path with existing scopes in the Library curatorial workflow tool. At this point it becomes possible to detect whether the issue is a crawl directive error or something more problematic in respect to the composition of the live site and rendering behaviors in use. The crawl vendor can be consulted to investigate the crawl logs to confirm a point of failure.

Results of capture assessment processing and subsequent investigative work are relayed back to the reviewers via email and are also included in comments within the Library's curatorial tool. These comments allow future stewards of the permanent collections to take stock of capture quality at a given time and collate known quality issues of a given seed.

VI. Conclusions

After implementing practical methods to satisfy each component of the grounded theory for web archives QA, the WAT has found that each practice provides a unique view into the quality of the web archive, with little overlap. After the first six months, it appears that staff performing capture assessment are reviewing captures not normally highlighted during the semi-automated size relevance assessments performed by WAT. This

indicates the importance of maintaining an ecosystem of quantitative and qualitative methods to assess quality, particularly as the collection continues to grow.

The emerging average correspondence score of 3.86 is a positive take away for the WAT. Results of web archiving at-scale can never be perfect, and this score indicates to us that captures are generally good. Correspondence ratings broken down by category are also positive indicators: 69% of captures scored a 4 or 5 on Completeness, about 64% scored 4 or 5 on Visual Correspondence, and 72% received a rating of 4 or 5 in Interactional Correspondence; only about 7% scored a 1 (lowest score) in any of the 3 correspondence categories. An anecdotal, positive takeaway of capture assessment is the WAT's ability to act in many cases to resolve or clarify "missing" elements.

VII. Ongoing Work

As the Library continues to work closely with its crawl vendor on QA, and particularly issues relating to archivability, the WAT is exploring other areas for improvement in the capture assessment and QA processes. There are some technical hurdles related to available tools for the workflow. WAT's first question in the capture assessment form, "is this the right website?" is meant to address the issue of link drift. If a capture is not intellectually consistent with the entity targeted for harvest, often this means that there is content drift on the live web. When checked "no", the form is supposed to end, however it defaults to all 5's (minor priority) and has affected 4 assessments out of the 193, at this point. This can be resolved by making the default ratings all "1" however this creates extra work for reviewers rating perfect captures, as they will have to manually click "5", "5", "5". Not having a default selection is not an option in the available tool.

Plans are underway to include employing technicians in the Library's Digital Content Management Section to complete capture assessments. As nominators have a small percentage of time for their web archiving duties, the technicians will be able to review a larger swath of the archive in a shorter time period. This practice will remove subject expertise, to some degree, but as they complete capture assessments, the technicians will gain familiarity with the collections. Data dashboards are also currently in development that can merge and visualize capture assessment results and technical crawl data (bytes, hops, etc.) for seed URLs and collections over time.

The Library's Web Archiving Program exists in a state of continual improvement, and the team will streamline features of the described workflows, as possible. Parts of the size relevance assessment workflow are scheduled to be automated further, such as generating the crawl report spreadsheet via continuous integration pipeline, thereby allowing WAT staff to press a button versus running a command line Python script. Against the scale of the archive, these small components of workflow preparation add up and the WAT will continue to leverage automation as much as possible.

1. REFERENCES

[1] Library of Congress Collection Policy Statements, https://www.loc.gov/acq/devpol/

[2] Bicho, G., Lyon, M. (May 24, 2022) Building a Sustainable Quality Assurance Lifecycle at the Library of Congress, presentation at the International Internet Preservation Consortium (IIPC) General Assembly and Web Archiving Conference. https://digital.library.unt.edu/ark:/67531/metadc1983138/

[3] Reyes Ayala, B. Correspondence as the primary measure of information quality for web archives: a human-centered grounded theory study. Int J Digit Libr 23, 19–31 (2022). https://doi.org/10.1007/s00799-021-00314-x

[4] Bragg, M., & Hanna, K. (2013). The web archiving life cycle model. https://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf

[5] Wappalyzer, https://www.wappalyzer.com/

[6] Seeds (seeds-report.txt), Heritrix, https://heritrix.readthedocs.io/en/latest/operating.html?highlight=seeds-report.txt#seeds-seeds-report-txt

[7] Bicho, G. (2021). The Library of Congress Web Archiving Team Goes Agile. The Signal: Digital Happenings at the Library of Congress. https://blogs.loc.gov/thesignal/2021/01/wat-goes-agile/

A Storage and Search Demonstration with DNA-Encoded Text

Laurel Provencher
Catalog Technologies
USA
laurelp@catalogdna.com

Swapnil Bhatia
Catalog Technologies
USA
sbhatia@catalogdna.com

Sean Mihm
Catalog Technologies
USA
sean@catalogdna.com

Abstract - DNA-based data storage (DDS) holds promise to deliver a paradigm shift for long-term, secure storage of data. To tap into this potential, methods must be developed to produce data-encoding DNA molecules with cost- and time-effective processes. Combinatorial synthesis of DNA molecules from prefabricated fragments of DNA offers a solution to this challenge. We are developing a DNA-based platform combining encoding algorithms, high-throughput synthesis, post-synthesis processing, sequencing, decoding algorithms, and DNA computing architectures into a unified system. DNA datasets encoding images and literary works have been successfully created and translated back into conventional data files containing the entire original set of data or a targeted subset of data. In this work, we demonstrate the ability to search for specific molecules encoding a specific word in a DNA dataset encoding the complete text of multiple literary works.

Keywords – DNA, sustainability, storage, search

Conference Topics – Sustainability, From Theory to Practice.

I. Introduction

DNA is the densest known information storage medium capable of supporting a diversity of operations including writing, reading, copying, and certain massively parallel models of computation. DNA is several orders of magnitude more resilient to natural degradation over time than other extant storage media, with a lifetime in the range of 1000s of years. It can be stored in a dry form requiring minimal space and little or no cooling. DNA is also amenable to a wide array of useful chemical methods that scale favorably in cost and energy requirements with the length and diversity of DNA sequences. Technologies for automating DNA synthesis, quantification, purification, sequencing, and chemistry have improved exponentially in capacity, performance, and cost in the past two decades [1].

These observations have led to the emerging field of synthetic DNA-based data storage and computing (DDSC) and the exploration of DNA as the information carrying medium underlying a digital data platform. When successfully implemented, our approach to DDSC will offer a novel option for archiving data at petabyte scales. Platform development will encompass strategies enabling energy efficient options for periodic information extraction and massively parallel computation.

We are seeking input from digital archiving professionals to learn how conventional archiving processes could be re-imagined with DDSC. As performance and scale of DDSC improves, questions around the design of the archiving ecosystem become more important. We encourage this community to help define the minimal requirements that must be met by a DDSC archiving solution.

II. Combinatorial Synthesis Strategy

Most approaches to encoding data into DNA rely on a direct translation between binary source alphabets and quaternary DNA alphabets. For example, “00” → “A”, “01 → “T”, etc. They require the synthesis of a completely new DNA polymer, base-by-base, to produce the molecular dataset. This is infeasible at scale without innovations addressing difficult chemistry and physics challenges.

We have developed a unique DNA data storage scheme which encodes data using a collection of disjoint sets S_0, S_1, S_{n-1}, each set containing distinct DNA molecules which we call components. Each

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

component in S_i is designed such that it can concatenate with any component in an adjacent layer. Together, the cartesian product of the sets S_0 x S_1 x … S_{n-1} defines a combinatorial space of DNA molecules ("identifiers") that can be constructed by concatenation of components. We impose an order on each component set and extend this to a lexicographic order on the combinatorial space. We may then treat this combinatorial space as a linear address space. To write a bit value of "1" at an address, we assemble the corresponding DNA identifier using its constituent components and to write a bit value of "0" we do not assemble the identifier corresponding to that address.

A primary advantage of our scheme is that writing relies on rapid self-assembly from a small, fixed set of components, a process amenable to fast self-assembly chemistry, parallelization, and high-throughput automation, rather than base-wise sequential synthesis. Given n component sets each of size c, the size of the combinatorial space defined increases exponentially with the number of layers (c^n) and multiplicatively with the number of components (c x n) with only additive increase in component library size. Thus, the approach is highly scalable.

III. Writing and Reading Data

To assemble identifiers from DNA components correctly, efficiently, and at high throughput, we have prototyped the Shannon system. This print engine contains an array of inkjet printheads to dispense any combination of up to 114 different DNA components as well as reagents necessary to covalently link components. A substrate is continuously fed underneath the printhead array and different combinations of components are overprinted into specific locations to create droplets containing unique sets of DNA fragments. The substrate moves from the printer array into an incubator chamber, then through a collecting mechanism which combines the droplets into a collection vessel [7].

Ordered assembly of identifiers from components is achieved through specific sequence design of the components and the intrinsic base-pairing behavior of DNA. As a material, DNA normally exists as a double-helix structure composed of two polymeric (chain-like) molecules twisted around each other. Each monomer (link of the chain) is one of the four possible nucleotides designated as A, T, C, or G. When the two polymers, or strands, wrap around each other to form a double helix, pairwise bonds form between the complementary nucleotides: A:T and C:G [3].

Each of the 114 DNA components we use to build identifiers are made of a central "barcoding" region of double-stranded DNA surrounded by single-stranded "overhang" regions. The nucleotides in the overhang regions are specifically designed to only complement DNA components from the appropriate adjoining layer of the assembly. When the overhang regions pair together perfectly, the ligase enzyme can create a covalent bond to permanently link the components.

The collection sample from the Shannon system, containing a highly diverse pool of assembled DNA molecules, is processed via a set of standard biochemical lab procedures to concentrate, isolate, and make copies of the successfully assembled DNA identifiers that encode each byte of data. This primary dataset can be split and stored in multiple locations as a liquid or dry sample.

A key tool for accurately reproducing either an entire DNA dataset or a targeted portion of a DNA dataset is the Polymerase Chain Reaction (PCR). This well-established technique uses heating/cooling cycles and a polymerase enzyme to disassociate the two strands of a DNA double-helix and build new strands of DNA complementing each of the separated strands. PCR can exponentially amplify specific DNA molecules in a sample by using specific paired 'primer' sequences in the reaction. Each primer is a short piece of single stranded DNA that complements a unique sequence in each of the separated strands of the molecules targeted for amplification. The polymerase will only make a copy of DNA if it finds a primed region to start building the complementary strand [4].

The DNA components used to create identifiers in our encoding scheme are organized in a hierarchical structure that allows replication of specific subsets of data via PCR. By performing one or more rounds of PCR with specifically designed primers, molecules representing specific elements of the data can be selectively amplified. This strategy allows us to access fractions of the dataset at different levels, as if accessing a specific file in a

directory composed of multiple layers of folders. Alternatively, we can target amplification of identifiers representing a specific word and its position in an ordered string of words.

Once a dataset or subset of the dataset has been specifically amplified via PCR, the DNA sequence of the identifiers can be read with established DNA sequencing platforms normally used for life-science applications [5]. Sequencing the DNA returns results indicating the order of A, T, C, and G nucleotides in a large random subset of DNA molecules present in the amplified sample. The DNA sequences are passed through a decoding algorithm to determine the word and its position in the full text file by determining the molecules' address in the combinatorial tree created during the initial encoding step.

IV. Proof of Concept

To demonstrate the capabilities of our platform, we encoded the complete text of eight of Shakespeare's tragedies, totaling 208,183 words. We then demonstrated a search over this DNA-encoded dataset for a specific query word. Importantly, the time and cost of our search strategy was independent of the size of the dataset. Our approach is founded on targeting and isolating identifiers corresponding to the specific query word. This approach is independent of the size of the dataset as all molecules are targeted in parallel in one step. Therefore, we expect our approach to scale with data size without a commensurate linear increase in the number of steps. We expect our approach will use fixed resources and steps when searching datasets containing up to 100s of millions of words of text.

V. Conclusions

In addition to benefits associated with data density and durability, DNA-encoded data holds potential to enable a new paradigm for performing parallel operations on large datasets. In our example of search and retrieval from encoded text, the efficiency of the operation is governed by the chemistry of molecular interactions. Unlike conventional computing, as the size of the dataset increases, time and energy required for the molecular interactions remains almost constant and enables larger datasets to be processed in the same amount of time.

To illustrate the fundamental differences in scaling a text search with molecular data vs. conventional data, we will use an analogy. Imagine that every word in a text file is represented by a fish. The fish is composed of multiple segments, some of which represent the position of the word in the file and some of which represent the actual word. In this example, the segment with the word of interest is identifiable because it is magnetic.

To perform a word search analogous to conventional computing, each fish must be individually examined to determine if it has a magnetic segment. Thus, one can imagine sending each fish through a narrow pipe and asking whether or not it adheres to a magnetic sensor. As the number of fish in the 'dataset' increases, so does the amount of time it takes to interrogate the complete population for the 'magnetic' word. The only way to speed up the search is to increase the number of pipes and sensors. Likewise, with conventional computing, the only way to scale data processing is to increase the number of individual processors to accommodate larger datasets.

A different way to identify all fish with a magnetic segment would be to drop one or more magnetic probes as 'fishing lines' into the pool. All magnetic fish will be attracted to the magnetic 'lures', with a speed that depends on the strength of the magnetic field and the physical distance between any individual lure and target fish. As non-magnetic fish will pose minimal interference, the total population of fish can scale considerably without significantly scaling the amount of time necessary to conduct the search. This fishing lure approach is similar to DNA-based computing where molecular interactions between complementary DNA strands behave analogously to nano-range highly selective programmable magnets.

To our knowledge, this is the first demonstration of a search mechanism working directly on raw molecular data. The ability to search data without first translating it from DNA back to conventional code means that the entire archive need not be read back into conventional computers. Rather, a selection process may be used so that the cost of sequencing the DNA data is only expended on the portions of interest. This consideration, together with the anticipated benefits of DNA as a compact and energy efficient solution for long-term data

storage, makes combinatorial DNA synthesis an exciting potential solution for digital archiving.

1. REFERENCES

[1] Dong Y, Sun F, Ping Z, Ouyang Q, Qian L. DNA storage: research landscape and future prospects. National Science Review. 2020.

[2] Bhatia, S. Turing Meets Watson-Crick: A Massive Data Storage Platform for Extreme Longevity, Density, and Replicability. Extremely Large Databases (XLDB) April 30- May 2 2018: Stanford, CA, USA.

[3] Ferry, G. The structure of DNA. Nature 575, 35-36. 2019.

[4] Smith, M. Polymerase Chain Reaction (PCR). National Human Genome Research Institute. 2023. https://www.genome.gov/genetics-glossary/Polymerase-Chain-Reaction

[5] Goodwin, S., McPherson J.D., McCombie W. R. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016.

[6] Bhatia, S, Gildea, K. A combinatorial writing scheme for high throughput DNA data storage. USENIX File and Storage Technologies. February 22-24, 2022: Santa Clara, CA, USA.

[7] Roquet N, Bhatia SP, Flickinger SA, Mihm S, Norsworthy MW, Leake D, Park H. DNA-based data storage via combinatorial assembly. 2021. bioRxiv doi: https://doi.org/10.1101/2021.04.20.440194

PDF/MAIL

Moving Theory Towards Practice

Tom Habing
University of Illinois Urbana-Champaign
USA
thabing@illinois.edu

Peter Wyatt
PDF Association
Australia
peter.wyatt@pdfa.org

Eden Irwin
University of Illinois Urbana-Champaign
USA
edeni2@illinois.edu

Ruby Martinez
University of Illinois Urbana-Champaign
USA
rubylm2@illinois.edu

Duff Johnson
PDF Association
Boston,
USA
duff.johnson@pdfa.org

Christopher Prom
University of Illinois Urbana-Champaign
USA
prom@illinois.edu

Abstract – Email is one of the most ubiquitous forms of communication in both personal and professional contexts. The EA-PDF (Email Archiving with PDF) project is developing a PDF specification (PDF/mail) for email archiving, as well as an open-source tool to convert emails to the new PDF format. By creating a new specification defining common understandings for archiving email, the project aims for PDF/mail to lower barriers to effective email preservation, meeting the needs of the archive and digital preservation community, while interfacing with companies that implement software. This includes building a community to support the project, developing the PDF specification itself, and creating a proof-of-concept tool to convert emails to PDF. With a standard and tools for converting and viewing emails, archivists and other professionals—particularly those who do not have access to other technologies supporting email preservation—will have a straightforward and cost-effective way of preserving emails for posterity.

Keywords – Email; File Format; Specification Development; Software; Metadata

Conference Topics – From Theory to Practice; We're All in this Together

I. INTRODUCTION

Email is one of the most ubiquitous forms of communication in both personal and professional contexts. Institutions around the world rely on email for all levels of day-to-day operations. Despite its widespread use and importance as a communication medium, email collections are not being accessioned, processed, or made available in archives at the rate one would expect or hope. Nor are end users able to effectively manage their own email archives.

To address this issue, the EA-PDF (Email Archiving with PDF) project is developing a PDF specification (PDF/mail) for email archiving, as well as an open-source tool to convert emails to the new PDF format. PDF technology is already widely used in archives and has many benefits, including ease of use and compatibility with a range of software and hardware.

By creating a new specification and defining common understandings for archiving email, the project aims for PDF/mail to lower barriers to effective email preservation, meeting the needs of the archive and digital preservation community, as well as end users of email software. With tools available for converting emails to PDF, archivists and other professionals—particularly those who do not have access to other technologies supporting email preservation—will have a straightforward and cost-effective way of preserving emails for posterity.

This paper provides an in-depth overview of the EA-PDF project and the development of PDF/mail to date. This includes building a community to support

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

the project, developing the PDF specification itself, and creating a proof-of-concept tool to convert emails to PDF. This project and PDF/mail will serve to to make email archiving more efficient and accessible, helping to guarantee that emails needed for business, legal, historical, cultural, or personal reasons are preserved before lost to the ether.

II. Community Development

The project's first goal is to draw together and advance a cohort of individuals from archives and the PDF community, in formal conversation with each other. Parties with functional expertise in digital preservation, PDF standards development, and PDF technical implementation can iteratively develop the specification. This work centers in the activities of a liaison working group (LWG) hosted by the nonprofit PDF Association. Co-chaired by EA-PDF Project Director Christopher Prom and PDF Association Chief Technology Officer Peter Wyatt, the group has met on a bi-weekly basis since November of 2021. This collaborative structure has been critical to supporting the project. It enables participants with different interests and objectives to work together towards a common goal of developing a specification for using PDF to package and represent email. It takes advantage of the formal specification design process supported by the PDF Association (a non-profit trade group dedicated to providing a vendor-neutral platform for developing open specifications and standards for PDF technology).[1]

The specification aims to provide a clear and comprehensive set of guidelines for the development of the PDF/mail container. The LWG identified several core archival attributes to be included therein, including defined structures for email data, metadata, and attachments. It also including a reference copy of the source MBOX or EML, for backwards compatibility and provenance tracing. These package attributes will ensure that email messages, folders, and accounts may be packaged into archive-ready PDF packages for preservation and reuse in a formal archives, while also providing end users with a turnkey solution.

By utilizing PDF/mail as a standard format for packaging email, this project aims to address the challenges associated with email archiving, such as the lack of simple and accessible email preservation solutions that can be easily adopted by institutions. According to a survey distributed to Illinois repositories, about 60% of respondents indicated that they have not collected any email collections (Martinez et al., 2023). They survey also found that many archives lacked necessary training, technology, and scale of email, were barriers to preserving email with non-PDF tools. PDF/Mail directly addresses these issues by providing a low-barrier solution and building on a nearly ubiquitous technology.

Accordingly, the project aims not only to formulate the standard, but also to develop tooling that demonstrates the proof of concept. It complements existing approaches, offering a pathway to produce PDF files that fully encode email message metadata, content, and structure. This will allow for downstream uses of the files, such as ingest into digital asset management or digital archives systems. Depending on local policies, the PDF files might become the preservation master, or serve as an access copy, complementing an MBOX master.

III. File Format Development and Design Considerations

The second project goal is to leverage the community for specification development. Building on the work completed in phase one of the project (EA-PDF Working Group, 2021), the LWG is developing a detailed technical description for the PDF/mail file forma. This leverages the general-purpose PDF/A (archival) file format to meet archival needs for preserving and providing access to both the visible content of email messages and to embedded message metadata. Files complying with the specification will be usable in today's PDF viewers, (what we term 'legacy' viewers), but software designed for viewing PDF/mail files will provide a richer navigational experience.

In this way, PDF/mail is similar to ZUGFeRD, Order-X, and Factur-X, all of which use the archival specification for PDF, ISO 19005 (PDF/A) as a foundation and leverage 3rd party standards to define additional domain-specific aspects. However, due to the requirement to preserve source and provenance metadata and email attachments, there is a heavy technical dependence on files embedded

[1] https://www.pdfa.org/about-us/

inside the PDF/mail file, which requires alignment with (minimally) ISO 19005-3:2012 PDF/A-3 or PDF/A-4f (ISO 19005-4:2020).

Like PDF/A, PDF/mail includes both file format requirements and a limited set of processor requirements. Due to the variety of possible use-cases, many requirements are expressed as "should" (strong recommendation) rather than "shall" (hard requirement). Like other PDF subset standards, PDF/mail does not define the precise appearance or algorithms that convert an email to PDF, nor does it prescribe content details.

PDF/mail profiles will have three main use cases:

1. A s single email in a single PDF (PDF/mail-1s).
2. Multiple emails in a single PDF, but without a hierarchical or folder-like structure, such as from an MBOX files from an entire account (PDF/mail-1m).
3. Container PDFs which contain one or more PDF/mail files, for example, preserving someone's entire email output with various folders, such as Sent, Inbox, Draft, etc. as well as any custom organization scheme. (PDF/mail-1c)

At the time of this paper's publication, November 2023, the PDF/mail 0.3 spec was under discussion in the LWG, to be shared more broadly within the PDF, digital preservation, and archives communities in spring 2024. The draft specification includes these primary features:

1. PDF/mail files shall be PDF/A-compliant
2. The standard will support metadata describing a corpus of email messages, at the document level of the PDF file, such as name of account holder. Defining these attributes is the responsibility of the archival community, and we are aligning them with the standards of other projects, such as EPADD+
3. At the message level, a set of common email Header Fields are formally categorized as Core Header Fields. The Core Header Fields shall always be present in the "message-level" XMP using Document Part Metadata in each PDF/mail-1s and PDF/mail-1m file. They will also be visually present in the page content of the EA-PDF file using text objects.
4. Where possible, metadata will be mapped to standard Dublin Core metadata fields, such dc:creator for from, and pdf:CreationDate for Date.
5. All PDF/mail-1s and PDF/mail-1m files shall embed the original source email data (e.g., EML, MBOX, OST/PST, etc.).
6. PDF/mail will support richly formatted email body formats such as HTML and RTF.
7. All email attachments shall be represented inside EA-PDF files, as embedded file streams
8. PDF/mail creation software may additionally decide to preserve assets referenced in the source email (e.g., images, SVG), including by fetching assets from the internet, when avaiable.
9. PDF/mail permits, but does not mandate, that actionable links in the source email must be link annotations in the output PDF/mail.
10. PDF/mail will allow for preservation of complex hierarchies of folders containing emails, such as Microsoft OST/PST, as represented by many email clients.
11. PDF/mail will support document structure and navigation features including Tagged PDF and Document Part Metadata (DPM).

IV. PDF/MAIL TOOL PILOT

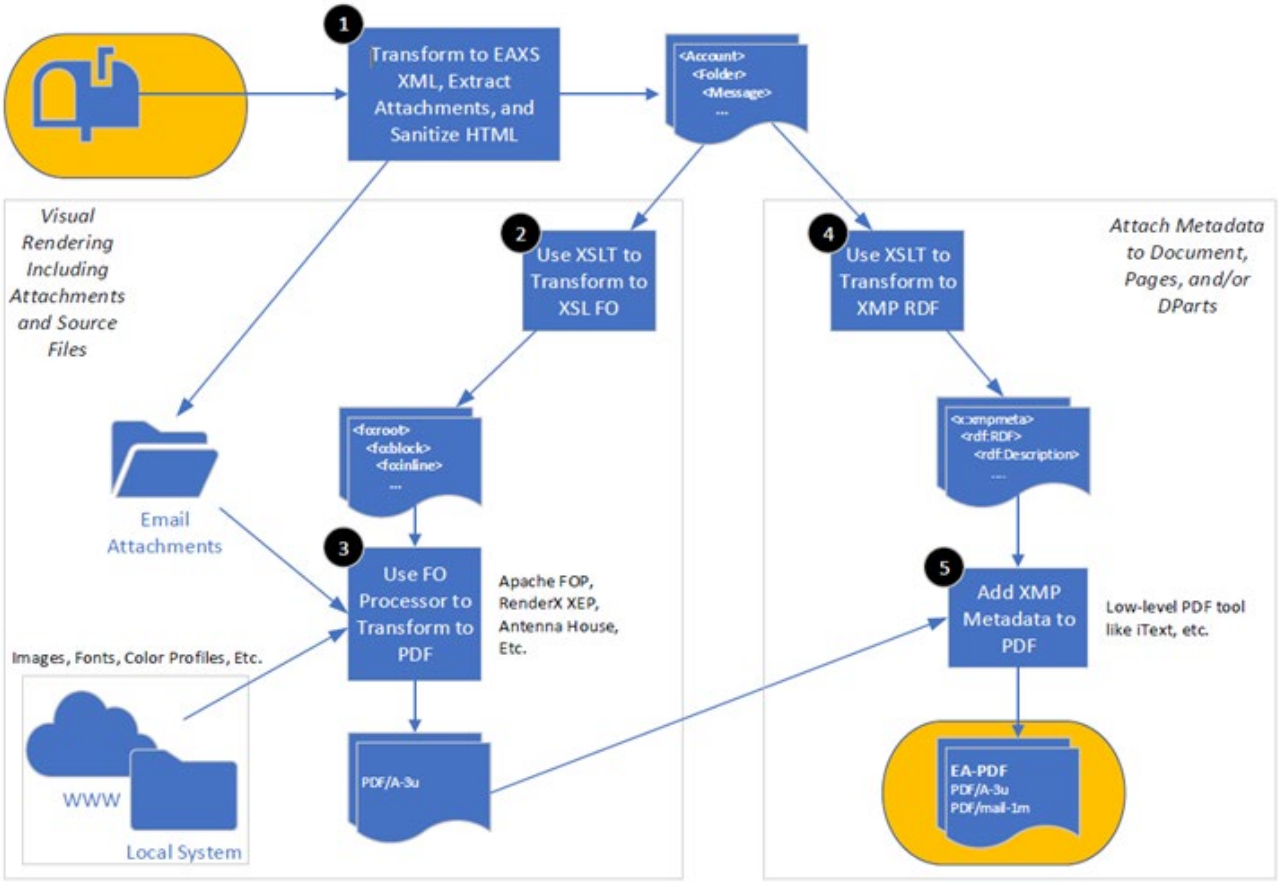
In parallel with the specification's development, the University of Illinois is developing an open-source PDF/mail creation tool. The parallel development of the tool has allowed participants in the LWG to react to draft outputs, and for the specification designer to incorporate feedback. Early versions of the code are available in our GitHub project, but at the time of publication, a distribution package has not (yet) been provided.[2]

[2] https://github.com/UIUCLibrary/ea-pdf

This section of the paper provides a high-level overview of that tool.

Primarily because of the developer's skillset, the tool was written in the C# programming language using the .NET Core cross-platform framework, allowing the application to be ported to Windows, macOS, or Linux. The tool also utilizes several other open-source libraries, including MimeKit[3] for parsing email mbox files, HtmlAgilityPack (HAP)[4] and Fizzler[5] for parsing HTML and CSS, Saxon HE[6] for XSLT transformations, Apache FOP[7] for converting XSL-FO into PDF, ItextSharp[8] for low-level PDF manipulation, among others. The GitHub project includes a basic command-line interface for the conversion tool. Referring to Figure 1, Process Flow for Conversion of MBOX to Archival PDF (EA-PDF), there are three major parts of the process: converting the email into XML, visually rendering the XML as PDF, and adding metadata to the PDF. A high-level description of these processes is below; numbers in the diagram match those in the text:

Figure 1. Process Flow for Conversation of MBOX to Archival PDF (EA-PDF)



1: Using custom code and the MimeKit, the mbox files are parsed and converted into XML files. The XML schema used for these files is a modified version of the EAXS schema[9] developed for the TOMES project. In addition to creating the XML files, this part of the process also extracts the attachments from the emails; these can be embedded in the XML as base64 encoded data or saved as external files. Any HTML message bodies are also cleaned up and converted to XHTML with inline CSS using HAP and Fizzler; this is done to accommodate the next transformation into XSL Formatting Objects (FO) [10].

[3] http://www.mimekit.net/
[4] https://html-agility-pack.net/
[5] https://www.nuget.org/packages/Fizzler.Systems.HtmlAgilityPack: /
[6] https://www.saxonica.com/welcome/welcome.xml
[7] https://xmlgraphics.apache.org/fop/
[8] https://github.com/VahidN/iTextSharp.LGPLv2.Core
[9] https://github.com/StateArchivesOfNorthCarolina/tomes-eaxs/blob/master/docs/documentation.md
[10] https://www.w3.org/Style/XSL/Overview.xml

2: Using custom XSLT, including a modified XHTML to FO transformation from Antenna House[11], the XML from step 1 is converted into XSL Formatting Objects (FO). The Saxon XSLT engine is used for the transformation. The XSL-FO is structured with a cover page, and each separate email message is started on a new page, along with a list of attachments at the end of the document. During this step Named Destinations are also added to the FO document for internal linking and so that metadata can be attached to the correct pages as described later in step 5. This step also affords some end-user customization; by modifying the XSLT, the resulting PDF rendering can be altered. It can also be customized to support different open-source or commercial FO rendering engines if desired.

3: Next, using Apache FOP or some other FO processor, the XSL-FO is transformed into PDF. Using processor-specific XSL-FO extensions or configuration settings, the PDF is made PDF/A compliant. The FO rendering engine also pulls the source mbox file and all email attachments into the PDF along with external resources from the web or local file system, such fonts, color profiles, or images linked in the HTML message bodies. This results in a PDF/A document with the significant email message headers rendered as readable text, the plain text and HTML messages bodies rendered as readable text, along with links to the embedded source file and attachments.

4: This step uses another custom XSLT to transform the EAXS from step 1 into XMP RDF metadata. A separate *rdf:Description* is created for each separate email message in the PDF. To the extent possible, predefined XMP properties[12] are used, but some custom properties have also been defined in an extension schema[13] for cases where there is not an equivalent pre-existing property, such as the email headers *to*, *cc*, *bcc*, *in-reply-to*, *references*, etc.

5: In this step, the XMP metadata created in step 4 is inserted into the PDF document and linked to the document or to the appropriate page or pages which correspond to the email message described by the metadata. The Document Part (DPart) Metadata (DPM) standard first introduced in the PDF/VT specification[14] is utilized for linking these metadata to the appropriate pages. DPart and DPM allow a set of pages, defined by a start and end page, to be associated with an XMP metadata stream. As mentioned, PDF Named Destinations inserted in the PDF during steps 2 and 3 are used to identify which pages represent which email messages. This step requires low-level manipulation of the internal PDF data structures; the open-source iTextSharp toolkit is currently used for this level of access. In addition to inserting message metadata, this step also inserts document-level metadata, primarily the XML extension schema describing our new non-standard metadata properties. This step can also perform other PDF enhancements that might not be possible using an XSL-FO processer alone (as described in steps 2 and 3), such as adding metadata properties to the attachments, adding watermarks, or setting the default PDF viewer settings like zoom level, etc.

At the end of step 5, the result is an archival PDF/A file which conforms to the new PDF/mail specification. Future enhancements to this tool might include a simple GUI interface for one or more platforms, improved customizations so that end-users can easily change the visual rendering of the PDFs or embellish the metadata with local customizations. Finally, follow-up work should include the development of tools that can render or consume the archival PDF/mail documents for use in a digital archive setting, such as user-friendly viewing of metadata, or extracting metadata for searching, categorizing, creating extracts, among many other archival functions.

[11] https://www.antennahouse.com/hubfs/uploads/XSL%20Sample/xhtml2fo.xsl

[12] https://www.pdfa.org/resource/technical-note-tn0008-predefined-xmp-properties-in-pdfa-1/

[13] https://www.pdfa.org/resource/technical-note-tn-0009-xmp-extension-schemas-in-pdfa-1/

[14] https://www.pdfa.org/wp-content/until2016_uploads/2011/08/Technical-Introduction-to-PDF-VT.pdf

V. Discussion

PDF/mail is a prospective, under-development solution to a known problem: The need for a simpler, easy-to-use email archiving and access format. Neither the specification nor the tooling described above are intended to offer the only or preferred method to achieve overall repository and institutional needs; digital preservation practice is too complex and varied to support normative solutions. PDF/mail has, in that respect, three goals.

First, PDF/mail is intended to provide the many institutions that have not previously engaged in archiving of email with a low-barrier method to do so.

Second, it provides those institutions and many others a distributable, access-forward format that can be accessioned, arranged, described and preserved within existing repository architectures, which often support PDF.

Finally—and this is perhaps a critical point—it aims to provide the community of those who develop software to support the file format with a clear statement of need and use cases to support end user needs (both personal and institutional), but in a way that also supports long term preservability.

In that sense, the PDF/mail proto-standard provides several opportunities for additional research, each of which deserve further exploration, extrapolation, and development.

As noted above, PDF/mail files will include rich, embedded metadata. Looking at this from an archivist's perspective, the PDF Dpart and associated Document Part Metadata reflect archival descriptive practices, which support both hierarchy and other forms of relationships. Document management systems, digital asset management systems, and digital library tools either include the ability to index and harvest embedded metadata or allow developers the ability to implement APIs and other tools to extract and index metadata on input. By providing metadata in a consistent, XMP and RDF-based format, the PDF/Mail standard seeks to enable indexing and discoverability of email messages alongside other digital content.

Similarly, the conversion of email content into PDF provides opportunities to better study, understand, and support end user archiving goals. Would it be helpful to provide users an easy way to archive their own personal emails to a format that does not require reimport to an email client, to make them useful? If so, what might an ideal viewing experience look like? What can archivists and industry professionals, working together, do to support it? And can that experience make it more likely that future generations will have access to the rich record of human experience and culture that is now locked away in the accounts of people famous, infamous, and everything in between?

VI. References

[1] EA-PDF Working Group. "A Specification for Using PDF to Package and Represent Email." Text. Board of Trustees of the University of Illinois, January 2021. https://www.ideals.illinois.edu/handle/2142/109251.

[2] Martinez, R., Prom, C., & Lee, C. (2023, January 13). "Practical" Vs. "Exemplary" Sustainability: Is There a Right Way to Archive Email?_20220914. https://doi.org/10.17605/OSF.IO/6JEBX

CREATING AN END-TO-END PROCESS FOR IMPLEMENTING A DIGITAL ARCHIVING WORKFLOW

How we are putting theory into practice

Leo Konstantelos
University of Glasgow
United Kingdom
Leo.Konstantelos@glasgow.ac.uk

Emma Yan
University of Glasgow
United Kingdom
Emma.Yan@glasgow.ac.uk

Abstract – This paper presents the efforts of Archives and Special Collections (ASC) at the University of Glasgow to produce an end-to-end process for implementing our digital archiving workflow. This will be achieved by conducting a pilot project and in this paper we discuss the project, our methodology and outputs and outcomes.

Keywords – Digital archiving, workflows, case study

Conference Topic – From theory to practice.

I. INTRODUCTION

Archives and Special Collections (ASC) in University of Glasgow Library is responsible for managing, promoting, providing access to and supporting engagement with the Library's unique and distinctive collections. These collections increasingly involve digital materials and in developing our digital archiving service we have begun to put a number of processes in place. In 2022 we used one of our collections, the NVA Archive, as a case study to assess our digital preservation capability [1]. After completing this study, we moved on to a new pilot project which is the focus of this paper.

The purpose of the end-to-end digital archiving pilot project purpose is to produce an end-to-end process, related procedures and methods for implementing the ASC digital archiving workflow [2]. The pilot has been allocated ten months and runs from December 2022 to September 2023, and the only required resource is dedicated staff time to develop and deliver the project.

II. METHODOLOGY

We started by outlining a project plan on what we wished to achieve with the pilot. We evaluated our existing practices against the workflow and identified where we need to extend and develop our services.

Within the ten-month period we intend to employ the following methodologies to develop, test and evaluate the necessary framework for delivering robust digital preservation and digital archiving services.

We intend on achieving this by:

- Developing an archival forensics workflow
- Extending our quarantine methods for storage media
- Creating a digital processing action logging system
- Creating an access procedure for the Archival Forensics Lab
- Identifying our digital media holdings
- Creating a digital holdings prioritization tool
- Conducting a collections development policy review
- Exploring appraisal and description with archival forensics
- Exploring transfer of research data at end-of-life
- Updating the risk register
- Reviewing donor agreements
- Implementing an end-to-end digital archiving case study

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

- Involving the DPC in the review and evaluation of developed procedures

The timescale for the work is planned on a Gantt chart, as follows in Fig. 1

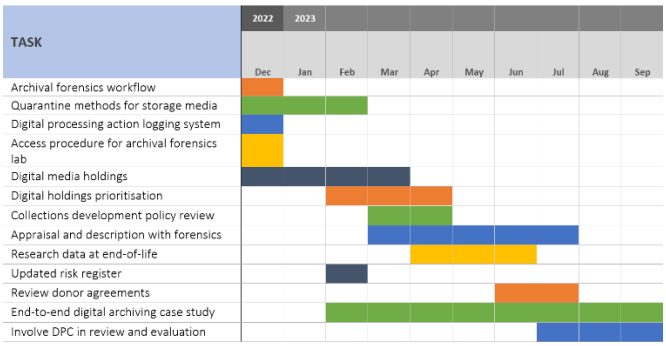


Fig. 1: Pilot project timeline

As seen above, we are starting with getting systems and procedures in place and most time will be spent on surveying our digital media holdings, appraising and describing collections using digital forensics software and on the end-to-end digital archiving case study.

Further detail on what we will achieve is as follows.

A. Archival forensics workflow

We have developed an archival forensics workflow for digital storage media [3] that operates as both standalone and as an integration with the digital archiving workflow. This workflow will be tested during the end-to-end digital archiving case study.

B. Quarantine methods for storage media

We will work with the ASC Conservation & Preservation team to extend the current quarantine procedures with actions catering for the specific needs of digital storage media. The intended output is a revised quarantine procedure and guidance.

C. Digital processing action logging system

We are exploring solutions to log digital archiving and preservation actions to records, in a manner that aligns with the archiving and forensics workflows. Unfortunately digital archiving and preservation actions cannot be logged in a meaningful way in our collection management system, so we are creating a separate database to log actions to collections where data can be transferred between the database and the system.

D. Access procedure for archival forensics lab

We are developing a procedure for access to the Archival Forensic Lab that provides conditions of access the Lab. Effective implementation of this policy will minimize unauthorized access to the AFL and further safeguard the collections.

E. Digital media holdings

We are surveying our collections and creating a list of physical storage media held by ASC across our University Archive, Business Archive, Theatre Archive and manuscript collections. Using this list we will identify the risk of loss based on medium type and condition.

F. Digital holdings prioritization

We have devised a methodology and tool for prioritizing the digital archiving and preservation processing of current digital holdings, using community resources and good practice guides.

G. Collections development policy review

In consultation with other members of the ASC team we will review and amend the Collections Development Policy to incorporate aspects of born-digital collections and methods of acquisition for digital records. This will include determining acceptable formats based on our current and projected digital archiving capabilities.

H. Appraisal and description with forensics

We are using our suite of digital forensics tools to forensically appraise digital acquisitions, following the archival forensics workflow. We have a Digital Intelligence FRED Forensic Workstation and FTK software and using these we intend to explore ways to leverage forensic technology to (semi-)automate metadata generation. We will focus on this during the end-to-end case study.

I. Research data at end-of-life

We will explore mechanisms and methods for transferring research data that have reached end-of-life and are considered archival records. This work will involve liaising with the Library's Research Information Management team who manage the digital research data repository.

J. Updated risk register

Work on the risk register has been ongoing since February 2021 and we intend to finalize updates to the risk register and prioritized risks list.

K. Review donor agreements

We will review and amend donor agreements to include clauses for forensic processing of digital acquisitions, especially regarding data carving for recovery of deleted files; decryption and password recovery for protected files.

L. Involve DPC in review and evaluation of developed procedures

We intend to engage with the DPC to help with reviewing and evaluating the developed processes and procedures.

M. End-to-end digital archiving case study

One major piece of work in this pilot is the case study, where we have selected two digital acquisitions to process and document from beginning to end of the digital archiving workflow, covering all steps and actions; and recording time taken to complete each. The case study will allow us to test a number of the newly created procedures discussed above.

Once the study is complete we will produce an end report recording our progress and decision-making, which we intend to publish.

III. About the Collections

We chose two collections to test during the pilot. Both are hybrid paper and digital collections and were chosen due to their size, complexity, content and perceived processing time as we want to ensure that we can complete the pilot project within the timeframe.

Our intent is to use these collections to test the workflow and while working though processing the collections, create policies, procedures and identify any sticking points or anything that is not working in the same way that we intended and make the necessary changes.

The first collection is the papers of Professor John Briggs, now an honorary research fellow at the University of Glasgow in the School of Geography and Earth Science, and previously Clerk of Senate and Vice-Principal (2012-2018) and Professor of Geography (1996-2012).

Briggs's research focuses on relationships between the use and management of natural resources and sustainable rural development in low-income countries, the impacts of structural adjustment policies in Africa on peri-urban development in the major cities and understanding the nature of agricultural landscapes in low-income countries. This collection, gifted to us in November 2021, focuses on his teaching materials as Professor of Geography, and the hybrid paper and born digital collection reflects the University's transformation from paper records to digital records, replacing print-outs and handwritten notes with PowerPoint presentations, Word documents and Excel spreadsheets. The digital material in this collection is fairly small in terms of size, amounting to 668MB, and was donated to us on one USB flash drive.

The second collection is the records of Dance House Glasgow, a creative arts organisation involved in supporting Glasgow's professional dance sector and offered community development programmes for over 20 years. In 2018 the company lost its Creative Scotland funding and ceased operating, and the collection was gifted to us via the Business Archives Surveying Officer in 2019.

The collection dates from c.1990 to 2018 and includes governance, financial, staff and project records, along with photographs, press cuttings, and promotional material. The digital material mainly consists of photographs and audio-visual material.

The physical extent of the digital material is 3 HDDs, 7 DVDs, 70 CD-Rs and 3 MiniDV cassette tapes. We know from our digital media holdings survey that one of the hard drives and 14 of the CDs were not working in January 2021, however we hope to interrogate this further using the archival forensic technology now at our disposal.

IV. Outcomes

The pilot is a work in progress and at the date of writing this paper as per our intended project timeline we have already completed some of the tasks outlined above.

The Archival Forensics Workflow is complete and has been published on COPTR's Community Owned Workflows webpage.

The access procedure for the Archival Forensics Lab is complete and is awaiting sign-off from ASC

senior management. We will offer a staff training session at to inform ASC staff of the procedure as well as the appropriate actions to take to request access to the Lab.

The storage media prioritization methodology and tool are complete, as are the updates to the risk register and the database for logging digital preservation actions.

We have started work on the end-to-end case study which has prompted us to start on using digital forensics to interrogate the files and test the archival forensics workflow.

We intend to share our methods and outcomes with the wider digital preservation community, and by the time of the iPres2023 conference in September we will be able to give a thorough account of the end-to-end digital archiving pilot project and how we translated the theory to practice.

V. REFERENCES

[1] L. Konstantelos, C. Paterson, E. Yan, "Evaluating Digital Preservation Capability With Large At-Risk Collections: Lessons Learnt From Preserving the NVA Archive," in iPres2022, Glasgow, 2022, pp.282-286. [Online]. Available: https://www.dpconline.org/docs/miscellaneous/events/2022-events/2791-ipres-2022-proceedings/file

[2] University of Glasgow Archives & Special Collections, Digital Archiving Workflow https://coptr.digipres.org/index.php/Workflow:Digital_archiving_workflow_(high-level)

[3] University of Glasgow Archives & Special Collections, Archival Forensics Workflow https://coptr.digipres.org/index.php/Workflow:Archival_Forensics_workflow_(storage_media_deposit)

I GOT A LETTER FROM MY PAST SELF

(Un)managed Change and Provenance

Rhiannon Bettivia
Simmons University
USA
bettivia@simmons.edu
0000-0003-4593-562X

Yi-Yun Cheng
Rutgers University
USA
yiyun.cheng@rutgers.edu
0000-0001-6123-7595

Michael R. Gryk
UCONN Health
USA
gryk@uchc.edu
0000-0002-3483-8384

Abstract – Significant properties (sigProps) research often focuses on the preservation targets. Yet research consistently shows that what is significant about an object is not necessarily inherent to objects. Simultaneously, sigProps research does not adequately attend to temporality. Time is built into the concept of sigProps: they are about what ideally should not change over time. This paper centers temporality in relation to sigProps to explore challenging case studies.

Keywords – provenance, managed change, identity, temporality

Conference Topics – Sustainability: Real and Imagined, From Theory to Practice

I. INTRODUCTION

Calvin: My past self is corresponding with my future self.

Hobbes: Too bad you can't write back.

--Watterson, 1995

Digital preservation recognizes that long-term preservation entails managed change. Managing change is necessary to ensure that users understand the overarching conceptual object as one and the same over time [18]. The need to imagine and plan for the future is one of the inherent challenges of digital preservation: digital preservationists must think like futurists [17]. Yet the relationship between identity and change is a quotidian concern. The cartoon character Calvin, of Watterson's Calvin and Hobbes series, constantly engages in time travel wherein he interacts with his future and past selves (Fig. 1). This comedic device points to the very real ways in which a person is, at different points in their life, both the same person and a fundamentally different person.

The challenges of identifying that which must change over time has impacts on digital preservation work across disciplinary spaces. In this short paper, we explore two research themes:

- Theme 1: In what ways is Past Calvin the same and different than Future Calvin?
- Theme 2: How do the nuances that distinguish people over time change when applied to physical and digital objects?

These themes have practical applications for digital preservation. Significant properties (sigProps) are "[t]he characteristics of an Information Object that must be maintained over time…" [9]. The concept of sigProps is both crucial and challenging: the need is acknowledged but the practice is hard. SigProps refer generally to the properties of a conceptual object that are required for its ability to establish its authority in the world. SigProps hinge on two key aspects: objects and time. In this paper, we focus on the temporal aspects and provenance in order to advance the scholarly conversation around the wicked problem of sigProps.

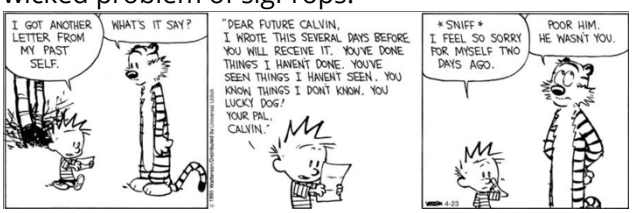


Fig. 1. Calvin and Hobbes, [April 20, 1995]

II. LITERATURE REVIEW

A. Temporal Provenance

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

SigProps support inherent change over time. Documenting these changes is part of telling the stories of objects, or provenance. Literature on temporal provenance focuses primarily on the e-sciences domains. Temporal provenance in scientific data is framed as (1) an ordered process based on causal relationships; (2) independent time slices; (3) circular processes.

Provenance models usually express time in an ordered fashion. For instance, in the Open Provenance Model (OPM), a second sequential process can only be initiated after a first process has occurred [15]. This suggests that the processes are directional, forming a directed acyclic, provenance graph.

In defining temporal provenance, Chen et al. discussed the potential of partial ordering of provenance graphs, and how one might be able to partition events into distinct time slices [5]. Similarly, Beheshti et al. proposed the Temporal Provenance Model (TPM) that puts time at the core in provenance documentation, as opposed to other event- or object-oriented provenance models [1]. In the TPM, time in provenance is captured not as a causal event, but as individual time-stamps to allow for versioning control of the same data objects. In this sense, time is an independent variable that partitions data objects into snapshots.

McPhillips et al. developed YesWorkflow, a scientific workflow management system built on the foundational concepts of retrospective and prospective provenance [13]. While retrospective provenance documents the execution, or past occurrences of a program, prospective provenance records the scripts, or the forward-looking recipes that enable a program to run. Here, the concepts of prospective and retrospective provenance are treated in a non-linear, circular fashion that supports a more nuanced approach to time in documentation.

Discussions of the temporal dimensions of provenance often center on metadata documentation, not on the data objects per se. Further investigation is needed on the use of temporal provenance to understand how data objects evolve over time.

B. Necessary Change

The Digital Preservation Coalition defines digital preservation as "…the series of managed activities…" [8]. In discussing artifactual objects, Owens [16] writes, "… what makes Mount Vernon *Mount Vernon*? Like all physical objects, it is changing at every moment" (p. 16). What are the sigProps of objects that are constantly changing? Historical contiguity is maintained through changes that comport with physical changes already happening: preservationists, digital or physical, roll with the changes that are going to come and make conservation decisions accordingly.

The question here, what makes a thing *that thing*, is central to the foundational understandings of the field of digital preservation. Thibodeau (2002) contributes terminological structure to the idea of the things that are the preservation targets: *that thing* is a conceptual object, supported by a pyramid of logical and physical objects. Preservationists make changes that can alter, re-order, substitute, or otherwise move the logical and physical pieces, while the top-level conceptual object must remain the same for the user in question. This approach mirrors models like the Functional Requirements for Bibliographic Records (FRBR), where the overarching conceptual work has various manifestations, expressions, and items that represent it [10]. The PREMIS metadata model also mirrors this structural approach to delineating *that thing* with its top-level intellectual entity object type [17].

Because of the foundational approaches digital preservation takes to *that thing* and managed change over time, it is a field that is poised to make broader impacts on issues at the intersection of the identity of objects and time. The following section employs case studies, biochemical research samples and video game franchises, to explore the themes stated at the outset.

III. CASES

A. Biochemical Research Samples

There is a renewed push to adopt persistent unique identifiers for samples in the natural sciences [4]. Biochemical samples are often altered, degraded or consumed in the process of a study, introducing the question of whether a persistent identifier is warranted for objects which themselves are not persistent.

In a biochemical laboratory, these ephemeral samples are typically given local identifiers, for instance with controlled experiments on multiple samples which vary in the concentration of a reagent or some other preparation step. This local identifier fulfills two simultaneous purposes: (1) it identifies the physical sample which is part of the experimental workflow and (2) it identifies the significant attributes of this particular sample with respect to the other samples which will be part of the study. In the latter case, a sigProp of the sample is its provenance - what it contains, how it was prepared, how it was treated, how it was stored, as well as temporal issues such as how long it has been since it was treated. Each of these concerns manifest itself on both the physical and concept level. It might be of importance whether a sample was stored at 4°C or at -20°. Alternatively, it might matter that a sample was stored in the 3rd floor freezer because there was a power outage in that room.

All of this is compounded by the fact that biochemical samples degrade over time. Samples age just as Calvin does, yet often on a timescale where the controlled variation between samples may be smaller than the variation within a single sample over time. This leads to some particularly tangled provenance stories when one wants to document the provenance of a sample and the methodology of an experiment in sufficient detail that it can be reproduced by others.

B. Super Mario

The previous case looked at the mechanics of organic change and the implications for identifying biochemical research samples over time. This section explores a socio-cultural example of the same phenomenon in the evolution of popular media figures over time, drawing from the work of McDonough and the Preserving Virtual Worlds grants [2,11,12]. *That thing* is Super Mario (Fig. 2), the Nintendo character who features in many media, starting with the *Donkey Kong* arcade games in the early 1980s.

The work of Preserving Virtual Worlds (PVWI and PVWII) is foundational to video games preservation. Two key findings that arose from PVWI are that (1) preserving interactive digital media requires a more systemic approach to determining sigProps even while acknowledging that (2) the preservation of popular games defies universal solutions.

Fig. 2 Uniqlo Super Mario 35th Anniversary T-Shirt depicting iterations of the character spanning the years 1985-2017, released in 2020.

PVWII identified the technical layers that make up a digital game as part of locating those sigProps. These layers include: the hardware/processor; the firmware; the software support; the physical; the application; and the experience layer [2].

Technological capabilities play a role in character design. Early design was frequently defined by the pixels and colors that fit within the storage and processing limits. Early Mario is pixelated in red, brown, and peach in 1988's Super Mario Bros. *(Mario 1)*. 2022's Mario + Rabbids Sparks of Hope is three-dimensional and brightly colored, wearing the iconic blue and red outfit (Fig. 3).

Fig 3. *Super Mario Bros.* (1985) and *Mario + Rabbids Sparks of Hope* (2022); images drawn from Wikipedia, image rights belong to Nintendo and Ubisoft.

At every layer of the technical stack, these versions of Mario are vastly different across a span of 37 years, including the processors, peripherals, displays, and experiences. Experiential differences are important, because this is where many users find the conceptual object in gaming. That it is possible to take the technological stack of the Switch and approximate the experience of *Mario 1* via Nintendo's emulator indicates that underlying

physical and logical pieces can change while the experience of *that thing* remains largely intact: this is a manifestation of sigProps in practice.

This case study is about the relationships between various manifestations of Mario (Fig. 2). Much as biochemical samples and Mount Vernon change over time, so has Mario over nearly four decades. When biochemical samples change in a lab context, the experiential differences might arise from their behavior in experiments. Marios differ in many ways over time. How and why do players recognize Mario as *Mario*? Part of the answer lies in how people make meaning of information. Clement traces how meaning is included in early information theories and she argues that users make meaning with information, rather than it being inherently meaning*ful* [6]. Marios remain *Mario* not just because of inherent characteristics like his blue and red costume, but because of meanings that come with interaction. The colors of Marios' costumes evoke a Mandela Effect: even when his outfit isn't actually red and blue, like in *Mario 1* or 1988's *Super Mario Bros. 3*[1], players remember *Mario* as red and blue.

McDonough notes that, "… [the p]reservation of computer games is in many ways a knowledge management problem, and without adequate metadata, managing the knowledge necessary to keep a game accessible and understandable is an insurmountable task." [14] This metadata is a form of provenance, and it must incorporate time: temporal framing for the objects and the temporal provenance that documents change in a way that enables objects to establish and maintain authority.

IV. Discussion and Conclusion

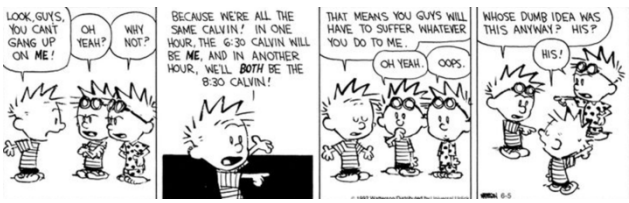


Fig. 4 Calvin and Hobbes, [June 2, 1992]

In a series of 1992 strips, Calvin attempts to avoid homework by time traveling to find a future Calvin who has already done it (Fig. 4). Unlike the arc where Calvin had a one-way conversation with himself via snail mail, here the Calvins literally find themselves in a room, communicating across time from 6:30-8:30, from homework time to bedtime. Ultimately, the 3 temporally differentiated Hobbes mediate the situation and do the homework. The aim of provenance documentation is to move beyond the one-way communication that comes from the past leaving missives for the future to something that resembles mediated conversations where past, present, and future can collaborate to form the best solutions. In previous work, we suggest that *subjunctive provenance* may improve provenance practice, acting as a mediator like the Hobbeses [3].

SigProps are inherently related to identity and time: they are the characteristics which determine whether the thing remains *that thing* over time. These cases demonstrate that significance is not necessarily inherent to an object: vastly different Marios are still experienced as *Mario*, the 3 Calvins are still just Calvin. Authenticity doesn't occur in a vacuum: meaning comes from experiences with objects rather than objects being inherently meaningful. Authenticity is a product of a relationship between objects and stakeholders [2,7].

The fundamental question remains: is the thing *that thing*? The answer is partly domain-dependent: in data management, it would be culturally common to see a change in a dataset resulting in a new data set, **Δ** dataset, even if the contents remained largely the same. However, a visibly obvious change in Mount Vernon, like the loss of a roof during a hurricane, does not result **Δ** Mount Vernon: it is still *Mount Vernon*. When Calvin tells himself, "You know things I don't know," he's talking about his own provenance: what differentiates the Calvins is what they've experienced. This raises the question: can provenance itself be employed as that which distinguishes a thing both as and from *that thing*?

These challenges are not academic. Practitioners manage diverse object and data types that behave differently enough that preservation and provenance practices are hard to universalize. Persistent identifiers that work for moon rocks do not work for biosamples. It is not that moon rocks don't change, but that the speed at which they do so is slower than a human life span, while biochemical

[1] The second image from the left in Fig. 2 is from this title, released in 1988 in Japan, 1990 in the US, and 1991 in Europe.

samples might change more through natural organic decay in a few days than they do in an experiment which is meant to alter them. Simultaneously, documentary processes that were done by hand for artifactual objects are impossible in computational environments: humans cannot document nanoseconds by hand. Incremental change is also a temporal facet that challenges documentary practices: there is a saying that it takes 7 years for every cell in the human body to be replaced with a new one. This saying points to three things: (1) that biological matter is always in a state of flux and change; (2) that humans assign symbolic meaning to this type of change; and (3) that humans understand incremental changes differently than other types of alteration. This type of biological incremental change is analogous to the Ship of Theseus story; it's the same kind of scenario that digital preservationists face when trying to track the knowledge base of a designated community.

This short paper presents a progressive idea: that digital preservation has not yet dealt sufficiently with the temporal aspects of sigProps. Time is always there in preservation work, but often at the periphery, where the changes of the object are documented and not the change of time itself. When that happens, difficult scenarios challenge existing models– Marios, Calvins, biochemical samples. This leads to a proliferation of standards and extensions, like the *provlets* of PROV, without solving the underlying issues. SigProps research often focuses on the preservation targets. Yet research consistently shows that what is significant about an object is not necessarily inherent to objects. Simultaneously, sigProps research does not adequately attend to temporality. Perhaps because time is part of the definition of sigProps, and part of digital preservation overall, it has been taken for granted and its role has been underexplored.

V. Acknowledgements

MRG acknowledges partial support from National Institutes of Health grants GM-111135 and GM-109046 and National Science Foundation grant 194670.

VI. References

[1] Beheshti, S., Motahari-Nezhad, H. & Benatallah, B. Temporal provenance model (TPM): model and query language. ArXiv Preprint ArXiv:1211.5009. (2012)

[2] Bettivia, R. Mapping Significance of Video Games in OAIS. *IPRES*. (2016)

[3] Bettivia, R. Cheng, Y-Y, & Gryk, M.R. What does provenance LACK: how retrospective and prospective met the subjunctive. In *Information for a Better World: Normality, Virtuality, Physicality, Inclusivity: 18th International Conference Proceedings*, 74–82 (2023)

[4] Buys, M., & Lehnert, K. (2021). Bringing together communities: IGSN and DataCite. DataCite. (2021). https://doi.org/10.5438/THHF-KX17

[5] Chen, P., Plale, B. & Aktas, M. Temporal representation for scientific data provenance. *2012 IEEE 8th International Conference On E-Science*, 1-8 (2012)

[6] Clement, T. The ear and the shunting yard: Meaning making as resonance in early information theory. *Information & Culture*. **49**, 401-426 (2014)

[7] Dappert, A. & Farquhar, A. Significance is in the eye of the stakeholder. *Research And Advanced Technology For Digital Libraries: 13th Euro- pean Conference, ECDL 2009, Corfu, Greece, September 27-October 2, 2009. Proceedings 13*, 297-308 (2009)

[8] Digital Preservation Coalition. Digital Preservation Handbook, 2nd edition. *Digital Preservation Coalition*. (2015)

[9] Grace, S., Knight, G. & Montague, L. Final Report. *InSPECT*. (2009)

[10] McDonough, J., Kirschenbaum, M., Reside, D., Fraistat, N. & Jerz, D. Twisty little passages almost all alike: Applying the FRBR model to a classic computer game. *Digital Humanities Quarterly*. **4** (2010)

[11] McDonough, J., Olendorf, R., Kirschenbaum, M., Kraus, K., Reside, D., Donahue, R., Phelps, A., Egert, C., Lowood, H. & Rojo, S. Preserving virtual worlds final report. (2010)

[12] McDonough, J., Senseney, M., Bettivia, Rhiannon, Evans, C., Kraus, K., Kirschenbaum, M., Donahue, R., Egert, C., Phelps, A., Decker, A., Lowood, H., Rojo, S. Preserving virtual worlds 2 final report. (2013)

[13] McPhillips, T., Song, T., Kolisnik, T., Aulenbach, S., Belhajjame, K., Bocinsky, K., Cao, Y., Chirigati, F., Dey, S., Freire, J. & Others YesWorkflow: a user-oriented, language-independent tool for recovering workflow information from scripts. ArXiv Preprint ArXiv:1502.02403. (2015)

[14] Montfort, N. & Bogost, I. Racing the beam: The Atari video computer system. (Mit Press, 2020)

[15] Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J. & Others The open provenance model core specification (v1. 1). *Future Generation Computer Systems*. 27, 743-756 (2011)

[16] Owens, T. The theory and craft of digital preservation. (Johns Hopkins University Press,2018)

[17] PREMIS Editorial Committee. PREMIS data dictionary for preservation metadata Version 3.0. (Library of Congress, 2015)

[18] Thibodeau, K. Overview of technological approaches to digital preservation and challenges in coming years. *The State Of Digital Preservation: An International Perspective*. 4-31 (2002)

KEY ELEMENTS OF A FILE FORMAT STRATEGY

The only bad file format is one that hasn't been documented.

Tyler Thorsted
Brigham Young University
United States
thorsted@byu.edu
0000-0003-0292-0962

Within the Digital Preservation Community there are many references to policies on file formats, acceptable file formats, preservation policies and strategies, risk matrices, and action plans. All have the intention of defining and describing file formats and guiding decisions on which formats to preserve how, and when. My team and I originally created a File Format Action Plan, which was later migrated from OneNote to Confluence and then included more strategic plans for hundreds of file formats. This paper explores which key elements should be included in an effective file format strategy and the different ways such data can be used by people and systems. What works for one institution may not work for another, and the work created by a larger institution may benefit those with smaller resources.

Keywords – File Formats, Documentation, Registry

Conference Topics – We're All in this Together; From Theory to Practice.

I. INTRODUCTION

Recently I attended a webinar entitled, *"Do unacceptable file formats exist?"*.[1] The chat during the webinar was most telling in how everyone views the topic of file formats within their organizations. I observed that Institutional polices and available resources end up driving or limiting most of the work in creating strategies. My response to the webinar question is this: "the only unacceptable file format is one that hasn't been documented."

II. THE PROBLEM AT HAND

As digital preservation professionals we understand the work we do is more than a backup.

"A backup is a short-term data recovery solution following loss or corruption and is fundamentally different to an electronic preservation archive." [2]

"Digital preservation combines policies, strategies and actions that ensure access to digital content over time."[3]

Ensuring access to digital content over time is a monumental task. The last few decades have seen a number of changes in the way we interact with our computers and devices. This has led to an explosion of software releases and just as quickly, that same software becoming obsolete. Recent trends in software subscriptions models keep digital preservation professionals working tirelessly to ensure this access.

Preserving a set of born-digital files from a previous decade can be daunting as format identification tools may not always be able to identify the format. The file format may not be documented anywhere on the modern web. It may take a bit of sleuthing to find samples in order to understand which specific software created the files.

While some file formats were designed to be easily understood, there are many binary and container formats which end up requiring qualified guesses on their origin and signature.

In one instance, I was documenting a proprietary format and I felt I had gathered enough samples to identify the header and which bytes indicated version. When I reached out to the developer to confirm, their response was, "Please, do not use any hex editor and do not try to analyze the binary data file." This type of attitude makes preservation and

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

access difficult for many many formats, increasing the risk in preserving.

In contrast, another format I researched was popular for a short time in the 1990's, often bundled with scanning software. It was a raster image format which faded off into obsolescence. Although the specifications were made public at the time, all links had rotted and were not available in the WayBack Machine. I was finally able to track down a developer and they were happy to share a copy of the specifications! [5]

Documenting old and new file formats reduces the risk of obsolescence, and if shared, reduces duplicated efforts.

III. Key Elements

Files stored in a repository all have unique attributes and history. The extension is not the only element dictating how these files are identified, migrated, or rendered. Below are some additional key elements that can be included in a file format strategy.

A. Identification

File formats should be identified using tools which look closer at a file beyond the extension. File Format Signatures can change over time. PRONOM PUID's are often used as the standard identifier, but there are many other tools which can be used.

B. History or brief description

Record a little background on the file format and its use at your institution. Include a current status of the software and its support by the developer.

C. Registries

There are many registries which you can refer to. Build on these for your institution specific needs.

D. Version information

Each version of software will create new versions of a file format. Knowing which versions of a file format are compatible with corresponding versions of software is important for proper rendering.

E. Specifications

If specifications for the file format exist, a reference to them should be included. If the specifications are unpublished or proprietary, details about research can be documented here.

F. Software to open/render

List which software can open and render the file format. Rendering matters. Not all software will open a file the same way. [4]

G. Software for migration

Software used for migration or normalization can be different than what is used to render. This element can also list software to avoid as it may cause unwanted changes. Include a decision tree for when a file is migrated.

H. Software to extract key properties

Detail which software can be used to extract key significant properties from the file format and their use.

I. Significant Properties (TechMD)

List which properties of the file format are important to extract? A TIFF may be an excellent raster image format to preserve, but if compressed with LZW, it may present a higher risk. List minimum set of required properties per institutional policy.

J. Risk

Risk assessments or preservation levels of support documents are useful tools for guiding strategy. [6]

K. Software to validate

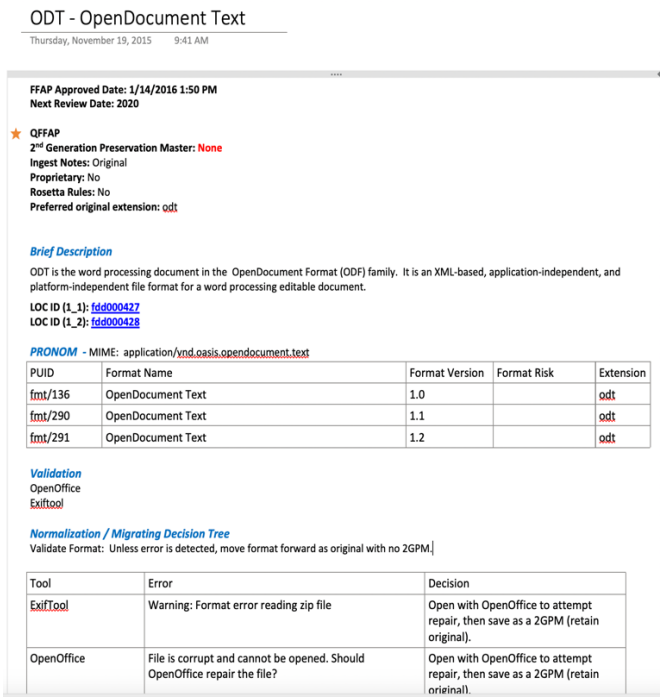
Many file formats can be validated to known specifications for institutional requirements. Software such as JHOVE or MediaConch can be listed here.

L. Rules

Many preservation systems have processing rules in place to help automate known identification and validation issues. Documenting these issues is important to understand decisions and preservation plans.

M. Platform (Mac/Win/Linux)

Some file formats and tools are platform-specific and require a certain environment to properly render or migrate.



ODT - OpenDocument Text

Thursday, November 19, 2015 9:41 AM

FFAP Approved Date: 1/14/2016 1:50 PM
Next Review Date: 2020

QFFAP
2nd Generation Preservation Master: None
Ingest Notes: Original
Proprietary: No
Rosetta Rules: No
Preferred original extension: odt

Brief Description

ODT is the word processing document in the OpenDocument Format (ODF) family. It is an XML-based, application-independent, and platform-independent file format for a word processing editable document.

LOC ID (1_1): fdd000427
LOC ID (1_2): fdd000428

PRONOM - MIME: application/vnd.oasis.opendocument.text

PUID	Format Name	Format Version	Format Risk	Extension
fmt/136	OpenDocument Text	1.0		odt
fmt/290	OpenDocument Text	1.1		odt
fmt/291	OpenDocument Text	1.2		odt

Validation

OpenOffice
Exiftool

Normalization / Migrating Decision Tree

Validate Format: Unless error is detected, move format forward as original with no 2GPM.

Tool	Error	Decision
ExifTool	Warning: Format error reading zip file	Open with OpenOffice to attempt repair, then save as a 2GPM (retain original).
OpenOffice	File is corrupt and cannot be opened. Should OpenOffice repair the file?	Open with OpenOffice to attempt repair, then save as a 2GPM (retain original).

Fig. #1, Example Strategy in Microsoft OneNote

IV. Audience

Who will be using this file format strategy? Is it just for preservation staff or is it intended for a broader audience? Institutional policies may be only useful internally, but documentation on file formats can be useful to share with the community.

V. Storing & Using the Data

Strategies can be documented in many ways. From simple Word Documents [7] to Excel spreadsheets [8], from Microsoft OneNote to Confluence. Others are using SQL databases or the popular Wikidata [9] ,Mediawiki approach. You can start small and grow the strategy over time or harvest from other sources into an actionable resource.

Digital Preservations Systems are moving toward more automated policies and preservation actions. These can be very useful, but don't let them replace your institutional strategies or be the only place such strategies are documented.

VI. Conclusion

Half the fun in documenting file formats is learning the history about the developer(s) and the purpose of each file format. Some were designed with the future in mind, while others were put together hastily to meet a deadline. Better still are the hidden meanings the developer left to be found by the curious (though, be careful of going down rabbit holes).

The statement, "The only bad file format is one that hasn't been documented" is not meant to convey that all documented file formats have no risk. It simply means that the more the community can document the formats in our repositories, the less risk they represent to preservation and access into the future.

1. REFERENCES

[1] A Panel Discussion: Do unacceptable file formats exist? February 9, 2023. http://bit.ly/3kVPNIn

[2] Digital Preservation: Continued access to authentic digital assets, JISC. http://bit.ly/3ystcGo

[3] "Definitions of Digital Preservation", American Library Association, January 18, 2010. http://www.ala.org/alcts/resources/preserv/2009def (Accessed March 8, 2023)

[4] Rendering Matters - Report on the results of research into digital object rendering, January 3, 2012 http://bit.ly/3kZUPUm

[5] XIFF File Format Research. http://bit.ly/3ZNZF5Q

[6] U-M Library's Digital Repository Services Registered Formats and Support Levels, http://bit.ly/3mAbqOY

[7] Strategies in Word Example. http://bit.ly/3F9fzA1

[8] NARA Risk Matrix in Excel. http://bit.ly/3Jotggl

[9] Wikidata as a digital preservation knowledgebase, http://bit.ly/3mALWBd

[10] Just Solve the File Format Problem Wiki, http://bit.ly/3myBhqp

NOTIONS OF VALUE IN DIGITAL OBJECTS

A debate with myself and others

Michael Popham
Digital Preservation Coalition
United Kingdom
michael.popham@dpconline.org
0000-0002-6842-4294

Abstract – The world of digital preservation and archiving has drawn heavily on the thinking of our analogue predecessors. When it comes to selecting materials, we are familiar with the idea of appraisal: "the process of determining whether records and other materials have permanent (archival) value" [1]. Typically, the notion of "value" is then further refined into broad sub-genres, such as evidential, informational, intrinsic, contextual, and so forth [2]. At iPres 2022, a panel session and related poster examined the problem of "The Value of Catastrophic Data Loss" but the debate repeatedly returned to measuring this value purely in terms of economic costs. This paper unpicks the notion of value further, and offers some reflections on how these ideas might apply to digital materials and be predicated on the essential differences between analog and digital sources.

Keywords – Appraisal, value, cost

Conference Topics – Theory to Practice.

I. INTRODUCTION (HEADING 1)

When it comes to collecting digital materials, appraisal is often one step in the accessioning process that nowadays is rather overlooked. Most of the collecting organizations and archives who present at iPres have very clear collection development policies and remits, and so need to give very little thought to the "value" of what they are collecting. They know what they need to collect, and why, and so can focus on that job and the associated challenges that arise from trying to preserve digital materials for any length of time.

Traditional archival practices have long ago settled on a consensus regarding the features of an analogue object that contribute to its intrinsic value (rather than its informational content) [3]. But discussions of digital materials can be somewhat reductive [4], and typically only discuss digital materials in terms of their value as surrogates for analogue items.

But surely there are some classes or aspects of digital objects that have a "value" that goes beyond the purely monetary (i.e. the economic costs of creation or replacement)? And in attempting to address this question I am conscious of the need to avoid straying into the intellectual weeds around notions of "significant properties" [5] and the like.

This paper attempts to explore some of the many ways that a digital object might be considered to be in some way "valuable", and implicitly suggests that the digital preservation community perhaps needs to broaden and update its thinking around the appraisal of digital objects.

II. INTRINSIC VALUE IN ARCHIVAL MATERIALS

It has been over forty years since the Archives Library Information Centre (ALIC) of the US National Archives published Staff Information Paper Number 21 on "Intrinsic Value in Archival Material" [3]. The paper states that "All record materials having intrinsic value possess one or more of the following specific qualities or characteristics" – and then goes on to list nine features of such records, namely:

1. *Physical form that may be the subject for study if the records provide meaningful documentation or significant examples of the form*
2. *Aesthetic or artistic quality*
3. *Unique or curious physical features*
4. *Age that provides a quality of uniqueness*
5. *Value for use in exhibits*

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

6. *Questionable authenticity, date, author, or other characteristic that is significant and ascertainable by physical examination*
7. *General and substantial public interest because of direct association with famous or historically significant people, places, things, issues, or events*
8. *Significance as documentation of the establishment or continuing legal basis of an agency or institution*
9. *Significance as documentation of the formulation of policy at the highest executive levels when the policy has significance and broad effect throughout or beyond the agency or institution*

The ALIC Paper then goes on to advise that records that have intrinsic value should be "retained in their original form if possible" and notes that "…opinions concerning whether records have intrinsic value may vary from archivist to archivist and from one generation of archives to another".

Whilst this document is clearly concerned with appraising analogue materials, can any of these qualities be reinterpreted and applied to assessing the ***intrinsic*** value of digital materials?

III. INTRINSIC VALUE IN DIGITAL MATERIALS?

If we take the first characteristic, "Physical form….", then whilst superficially this might seem irrelevant when we come to consider digital materials, surely the resurgent interest in emulation as a preservation strategy and the growth in computer museums, implies that there is something about the "original" form / appearance / rendition of certain digital materials that archivists recognize and value? This is particularly notable in the preservation of video games and early interactive works, where reproducing the look-and-feel of the material when it was first released is considered essential. Moreover, stories of retro games on their original (preferably untouched) media commanding eye-watering prices at auction are now commonplace [6].

One might argue that some of these same properties spill-over into the second characteristic of "aesthetic or artistic quality". This is best evidenced by the work of archivists involved in the preservation of digital works of art, who nowadays seek to work with artists to improve the likelihood that their creations will remain accessible to future generations. In addition, the furor around the prices paid for NFT artworks over recent years [7] arguably demonstrates that there are many people who clearly consider these digital materials to have intrinsic value – both aesthetic and economic.

It is perhaps less immediately obvious how digital materials might possess "unique or curious physical features" that attribute intrinsic value (item 3 in the ALIC list). Indeed, digital materials that are unique (or "curious"!) are likely to be very difficult to preserve, and so it seems implausible that such a characteristic would be seen in a positive light. One conceivable exception might be the case of program source code which includes the first use of particular algorithm.

I would suggest that for digital materials the characteristic of "age that provides a quality of uniqueness" (item 4 in the list), is still an emergent property. The commonplace digital preservation practices of data normalization and migration would seem to suggest that, as a profession, digital archivists rarely value the age of digital material per se (and even in those instances where an object is also kept in its original deposited form, this is primarily done as a safeguard against possible migration errors or as an indicator of provenance or authenticity, rather than because the original is valued for its age). However, the growing interest in historical computing, will surely lead to more digital objects being seen as having intrinsic value because of their age (e.g. early program code written in a particular language) – but such instances will surely be relatively few.

In contrast, it is relatively straightforward to think of instances of digital materials that will have "value for use in exhibits" (item 5). Whist the overwhelming majority of digital materials we collect and preserve may not display this characteristic, there are plenty of examples in existence – such as site CERN has created to recount "The birth of the Web" [8].

The sixth suggested characteristic indicating intrinsic value, "Questionable authenticity…." Is perhaps less likely to apply to digital materials. Appropriate metadata collected at the point of ingest, or the application of digital forensic techniques to the materials concerned, seem the most likely options to resolve any questions about authenticity or provenance. Failing that, computational analysis of the content of the material (e.g. stylistic analysis of an electronic document) may be sufficient to resolve concerns about its authenticity, in much the same way as we might use

handwriting analysis to discover the authorship of a manuscript.

"General and substantial public interest because of direct association…" (item 7 in the ALIC list) seems eminently likely to apply to digital materials as much as analogue. Digital archivists go to great lengths to preserve the provenance, authenticity, and integrity of the digital materials they collect, and so when they have records which pertain to a particular person, event, or issue, the association (and any concomitant suggestion of "value") can usually be demonstrated. When The Telegraph newspaper in the UK recently began publishing extracts from 100,000 WhatsApp messages sent by a former government minister during the Covid-19 crisis [9], despite the fact that those messages had not been properly collected, curated, and preserved, there was apparently no doubt in the public's mind that the messages were genuine. Even the ex-Minister concerned did not attempt to dispute the veracity of the messages, but rather took exception to his words being taken out of context – and encouraged his critics to read the complete exchanges before levelling their complaints. This would appear to be an area where the value of a collection of digital materials – certainly when expressed in terms of their utility – far exceeds what we might have expected from analogue counterparts.

The intrinsic value accruing from a digital record's "Significance as documentation of the establishment or continuing legal basis of an agency or institution" (item 8) seems to be self-evident. As new legal documents, agreements, and charters increasingly exist (only in) digital form, and as key players continue to digitize their analogue holdings of such records, then their intrinsic value seems to be widely accepted.

Likewise, digital materials "significance as documentation of the formulation of policy…" (item 9), with the ability of archivists to capture and record the fine details of a digital record's provenance, history of creation and updates, links to other digital materials, and so forth, would seem ample demonstration of their potential to possess this quality of intrinsic value.

IV. What's different about digital materials?

Having established that digital archival materials can satisfy many, indeed most, of the criteria used as possible indicators of intrinsic value, perhaps the obvious question that remains is: are there other qualities and characteristics that digital materials might possess in addition to those that have been suggested for analogue materials?

Perhaps one of the most obvious differences between archival materials in analog and digital form is that the latter typically offer greater utility. Compared to its physical counterpart, a digital record is often quicker and easier to create, reproduce, and share. It can be more readily stored, shared, accessed, and analyzed by tools which support a range of research activities. One might argue, therefore, that the greater the speed and ease with which digital material can be used, the greater its value to users – and so perhaps it follows that digital materials which conform to accepted and well-supported open standards and which are more amenable to study are inherently more 'valuable' than those which do not.

In a similar vein, the fact that most digital materials carry with them technical and sometimes descriptive information (e.g. in associated metadata), sometimes also details of how (and by whom) they may have been altered, and other evidence of their provenance and authenticity – all of which can be accessed and exploited relatively easily – might be said to enhance their value. The tools to unpick the history of a digital file are readily available to most digital archivists, whereas undertaking comparable research with analog sources often requires specialist skills and knowledge that is only available to comparatively few.

In crude terms it is also often far easier to establish and track the economic costs of creating, storing, and using digital materials than it is with analog records. Digital archivists typically have the information and tools to record the costs associated with born-digital or digitized materials, whereas comparable information about analog materials is often completely lacking.

Yet some of these very qualities which differentiate, and potentially add value to, digital materials as opposed to their analog equivalents, might arguably be said to *reduce* their value.

Whilst easy reproducibility is a helpful characteristic of digital materials, intuitively we feel that this reduces the sense of "rarity" and

"specialness" in such items, and this affects our judgement of their intrinsic value. Whilst a given digital object might be undeniably unique, the fact that one can create an absolutely identical copy with a few keystrokes influences our judgement of its worth not least because we know that we could not undertake the same action with an analog source. Even the very best facsimile copy of an analog source is never judged to be of equivalent intrinsic value as the original item from which the copy was made. Indeed, the value-laden terminology of "original" and "copy" seems to be considered largely redundant when we are examining digital materials – where there might be no way of distinguishing between two seemingly identical files.

Earlier, I asserted that many (indeed most) digital materials nowadays carry buried within them the metadata and other pieces of information necessary to establish their provenance and authenticity, and further suggested that this ascribed additional intrinsic value to a digital source. However, the very plasticity of digital information makes it all too easy to create, manipulate, or fake such details. From early examples of crude PhotoShopping, to the sophisticated deepfakes littering the internet of today [10], we have well-and-truly put paid to the adage that "the camera never lies" and have learned that we should no longer immediately trust what we can see with our own eyes. Fixity checks can help digital archivists identify any changes to digital materials that are in their care but they will not establish the veracity of the digital material itself. Likewise, although the use of digital rights management and digital signatures offer some degree of reassurance as to the provenance of material, most digital objects are not secured in that way. Those that are protected using such methods are most usually afforded this defense because of the perceived monetary value they represent.

V. Is it really all about the money?

I have made several assertions above that perhaps suggest the main characteristic that distinguishes digital from analogue archival materials, is their explicit or implicit economic value. Digital materials can be expensive to create, store, and manage – and anything which affects their usability, utility, or results in their loss, can be measured in cold, hard cash. Ransomware attacks are big news [11] and typically work by denying legitimate users access to data, rather than by removing or destroying the data itself.

Yet it might be argued that ransomware attacks or instances of data loss or destruction, do not actually alter the intrinsic value of digital materials in the same way that comparable incidents might affect analog materials. If a criminal were to burn the Mona Lisa, that work of art would be lost forever, but if a ransomware gang was to encrypt a company's essential data the company could largely mitigate the consequences of such a crime by keeping comprehensive, up-to-date backups that are (in every sense that matters), indistinguishable from the original digital records. So whilst it might be possible to put an economic cost on the data loss that results from the crime, this is perhaps better characterized as the costs of (temporary) loss of the utility and functionality made possible by the digital data, as with a suitable digital preservation strategy a bit-for-bit, byte-for-byte identical copy of the original material can be made available; something that could never happen with the smoldering ashes of the Mona Lisa, however technically proficient one might be.

I began this paper by stating that the focus of discussion would be on the non-monetary/economic value of digital materials, but I acknowledge that establishing value in such terms is not without its problems. Whilst we might be able to establish evidence for the financial 'input' costs of creation or replacement (e.g. how much it might cost to repeat the digitization of a particular manuscript, if an earlier set of image files were found to be unreadable or unavailable for some reason), there are many kinds of digital record which are literally irreplaceable (e.g. the astronomical data gathered from observing a comet which subsequently crashed into the sun).

Previous discussions about assigning monetary value to digital archives, such as Jeremy Heil's paper delivered to the Association of Canadian Archivists in 2017 [12], have explored the challenge of trying to establish a "fair market value" (e.g. for insurance or tax purposes) when there is no obvious market, or direct comparators, for a given set of digital records or material. And whilst establishing provenance of digital records might in many instances be easier and more reliable than doing so for their analog counterparts, it is less clear how conventional

notions like "condition" or original vs copy, might apply in a digital context.

Freda Matassa's book *Valuing Your Collection* [13], is a monograph entirely devoted to notions of "value" and how this term might be interpreted and applied to materials. Matassa notes:

> *The word 'value' has many meanings: price, worth, cost, significance, desirability, importance, asset, quality or excellence. It applies equally to financial or cultural worth. Curators know their collections in terms of significance. Stakeholders, however, often think of value only in financial terms. There may be times when both meanings coincide…* (*ibid*., p17)

The vast majority of examples discussed in the book relate to valuing analog materials (most often from museum/gallery collections), but there are some remarks which might apply equally well to digital materials in the context of this discussion, such as "For historic, scientific or aesthetic items, i.e. most of the objects in our collections, value cannot be based on the actual costs of production", (*ibid*., p19), and "Some items have very little or no market value, but considerable significance in the information they reveal" (*ibid*., p29).

When discussing factors which can influence the perceived (monetary) value of a work of art, Matassa makes an observation about authenticity which will ring particularly true with the community of digital preservationists, namely "Authenticity can make an item priceless and lack of it can equally render it worthless if it is found not to be by the artist or maker in question" (*ibid.,* p33). Digital archivists and preservationists have long recognized the importance of recording provenance, and using checksums to establish that something is unaltered, and one might almost be tempted to suggest that perhaps the "true" value (whether monetary or cultural) lies in the metadata of a digital object, rather than in the object itself.

VI. Conclusion

As may be all too evident, this paper is very much a thought-piece – and, worse, one without any clear conclusion. To me, it feels overly reductive and simplistic to measure the intrinsic value of digital materials merely(!) in terms of the monetary value that they represent simply because this can be measured (or guesstimated) using details we simply do not have about most analog archival materials. Moreover, the impossibility of retrospectively establishing a "fair market value" for many digital materials adds to the difficulty of trying to assign a monetary value to them once they have been created, collected, and ingested into a digital collection.

Digital archival materials offer so much more, in so many ways, than their analog antecedents. To ignore these characteristics and qualities when we talk about their value, or the consequences of their loss, seems to overlook the very features which make digital materials so valuable and worth preserving.

VII. Afterword

Any ideas or views expressed in this paper are entirely my own, and should not be attributed to my employer, the Digital Preservation Coalition. I have not shared or discussed these ideas with my colleagues, and as such I take full responsibility for any mistakes, omissions, half-formed statements, or wild assertions made herein. I offer these thoughts in the spirit of open debate.

1. REFERENCES

[1] *Archival Appraisal*. 26th June 2023, Wikipedia Contributors. Available at https://en.wikipedia.org/wiki/Archival_appraisal (last accessed June 26, 2023)

[2] *Exploring archival value: an axiological approach* Elaine Penn (Unpublished Doctoral thesis, University College London, 2014). Available at https://discovery.ucl.ac.uk/id/eprint/1455310 (last accessed June 26, 2023)

[3] *Intrinsic Value in Archival Material, Staff Information Paper Number 21*, Archives Library Information Center (ALIC), National Archives, 1982. Available at https://www.archives.gov/research/alic/reference/archives-resources/archival-material-intrinsic-value.html#note (last accessed June 26, 2023)

[4] *Digital artifactual value*, 26th June 2023, Wikipedia Contributors. Available at

https://www.wikiwand.com/en/Digital_artifactual_value (last accessed 26th June, 2023)
[5] *Significant Significant Properties*, R. van Veenendaal, P.C.M. Lucker, C.D. Sijtsma (2018). Available at https://openpreservation.org/wp-content/uploads/2018/10/Significant-Significant-Properties.pdf (last accessed 27th June, 2023)
[6] *22 of the rarest and most expensive big box PC games*, Ted Litchfield, March 13, 2022. Available at https://www.pcgamer.com/rarest-most-expensive-pc-games/ (last accessed 27th June, 2023)
[7] *Here Are the 10 Most Expensive NFT Artworks, From Beeple's $69 Million Opus to an 18-Year-Old's $500,000 Vampire Queen*, Sara Cascone, March 23, 2021. Available at https://news.artnet.com/market/most-expensive-nfts-1952597 (last accessed 27th June, 2023)
[8] *The birth of the Web*, CERN, 2023. Available at https://www.home.cern/science/computing/birth-web (last visited 27th June, 2023)
[9] *The Lockdown Files*, The Telegraph. Available at https://www.telegraph.co.uk/news/lockdown-files/ (last accessed 27th June, 2023)
[10] *What are deepfakes – and can you spot them?*, Ian Sample, 13 January 2020. Available at https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them (last accessed 27th June, 2023)
[11] *Ransomeware*, Wired, tagged articles of various dates. Available at https://www.wired.com/tag/ransomware/ (last accessed 27th June, 2023)
[12] Jeremy M. Heil, *Challenges in the Monetary Appraisal of Digital Archives*, presented at the Association of Canadian Archivists conference, Archives disrupted, Ottawa, ON, 8 June 2017. Available at https://qspace.library.queensu.ca/bitstream/handle/1974/26306/JeremyMHeil-MonetaryAppraisalDigiRec2017.pdf?sequence=3 (last accessed 27th June, 2023)
[13] Freda Matassa, *Valuing Your Collection: A practical guide for museums, libraries and archives*, (Facet Publishing, 2017), p.17 (electronic version, last accessed 27th June, 2023)

EVOLUTION OF BORN-DIGITAL MOVING IMAGE PROCESSING

Moving to scalable and sustainable workflows

Rachel Curtis
Library of Congress
USA
rcur@loc.gov

Laura Drake Davis
Library of Congress
USA
ladavis@loc.gov
ORCID 0000-0001-9892-2932

Abstract – Long-term preservation of born-digital moving image content is similar to that of any other file-based content in many ways. However, large file sizes, specialized equipment and resources, significant processing storage needs, and the movement of large files are challenges to creating sustainable and scalable workflows. The Moving Image Section of the National Audio Visual Conservation Center at the Library of Congress is making great strides in the development of sustainable and scalable workflows through an understanding of the technical infrastructure, moving image file characteristics and requirements, and the adoption of automated workflows using a combination of open source software and hardware resources.

Keywords – moving image, digital workflows, scalability, technical infrastructure, digital preservation

Conference Topics – From Theory to Practice; Immersive Information.

I. INTRODUCTION

The National Audio Visual Conservation Center (NAVCC) at the Library of Congress (the Library) is home to the world's largest collection of moving image and audio materials. NAVCC, as at other institutions, is experiencing a shift from analog to born-digital, and participates in broader efforts at the Library to establish a community of practice. However, the characteristics of born-digital moving image files present challenges in terms of file size, processing resources, storage allocations and network bandwidth. Over the last ten years, NAVCC staff have worked to address these challenges and anticipate future needs for born-digital moving image processing.

This paper discusses the evolution of born-digital moving image processing workflows, the impact of ongoing IT modernization efforts, resulting challenges in adapting to new internal requirements, and the efforts to ensure workflow sustainability when met with increased numbers of born-digital files.

II. INITIAL BORN-DIGITAL MOVING IMAGE PROCESSING EFFORTS

The Moving Image Section's first born-digital moving image collection workflow was developed for The HistoryMakers Collection. This significant collection consists of oral histories of prominent African Americans from a wide range of disciplines. The ingest of born-digital files was a new endeavor for the NAVCC, but a workflow was adapted from the existing digitization workflow, including verifying checksums, generating derivative files, creating ingest documents, and linking files to their corresponding metadata records.

As illustrated in Figure 1, the initial processing workflow for this project consisted of two parallel paths that converged before ingest. The NAVCC Moving Image Processing Unit created local

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

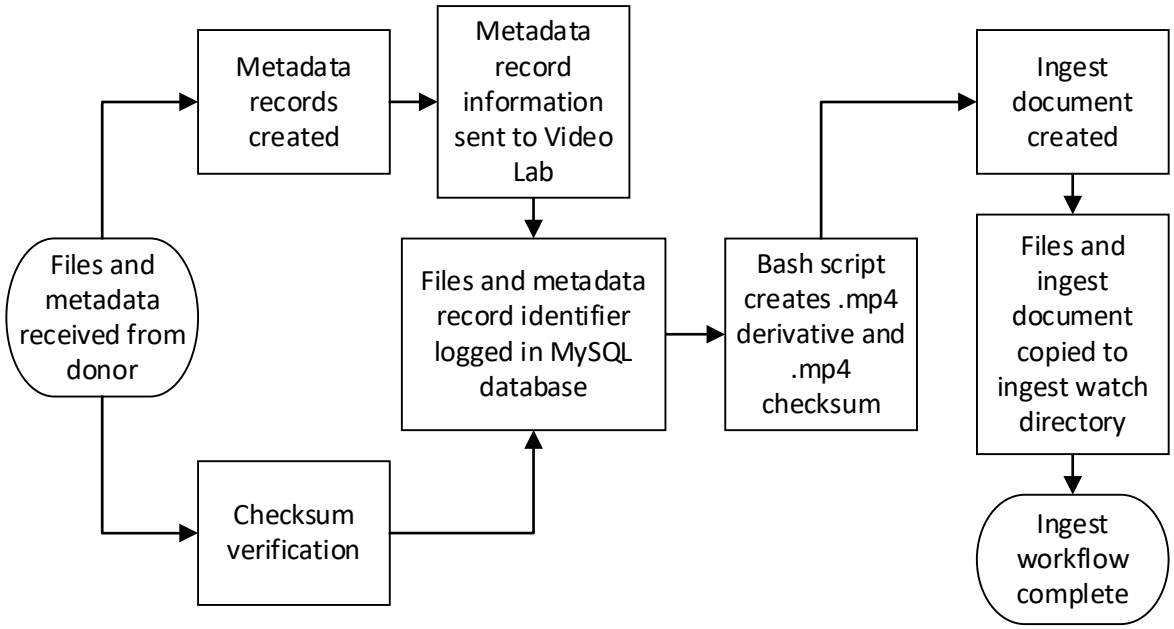


Figure 1: Initial born-digital moving image processing workflow

metadata records and Library of Congress Name Authority records; and the Video Lab created derivative files (.mp4) using OpenCube software [1] and ingested the preservation and derivative files. The metadata creation processes benefited from extensive information from The HistoryMakers organization including interview dates and interviewee biographical information.

This project demonstrated the need for staff dedicated to processing born-digital collections to ensure sustainability and the scalability. The initial workflow relied on the availability of the head of the Video Lab initiate and monitor the modified digitized workflow while balancing day-to-day Video Lab responsibilities. However, it would not be until the establishment of the American Archive of Public Broadcasting (AAPB) that a digital project specialist was hired and dedicated to born-digital collections.

III. A Born-Digital Program Begins

The AAPB began as a project funded by the Corporation for Public Broadcasting (CPB). In 2010, CPB conducted an inventory project and provided funds for 100 public television and radio stations to digitize items in their collection, which resulted in the creation of about 73,000 files. In 2012, CPB selected the Library and the public media station GBH to be the co-stewards of the archive. In this collaborative partnership, the Library is the preservation arm of the archive, ingesting high-resolution preservation files and ensuring their preservation for generations to come, while GBH makes files accessible on the AAPB website. In 2013, the Library hired a limited-term digital project specialist assigned to the AAPB with CPB funding.

The Library received 73,000 files on LTO tape in 2015. The files were delivered according to the BagIt specification [2] along with a master spreadsheet that contained filenames, metadata, and LTO barcodes after the vendor completed digitization. To facilitate the immense job of ingesting these files quickly, NAVCC staff adapted and expanded the scripts developed for HistoryMakers.

This initial AAPB workflow consisted of the following steps:

- Create a SQL database to store all datapoints
- Verify checksums in the bags
- Create a metadata record in the Library's MAVIS system for each file
- Move the media file and any sidecar files (such as .srt files) from the bags to a watch folder
- Create the ingest package for each set of files
- Ingest the files

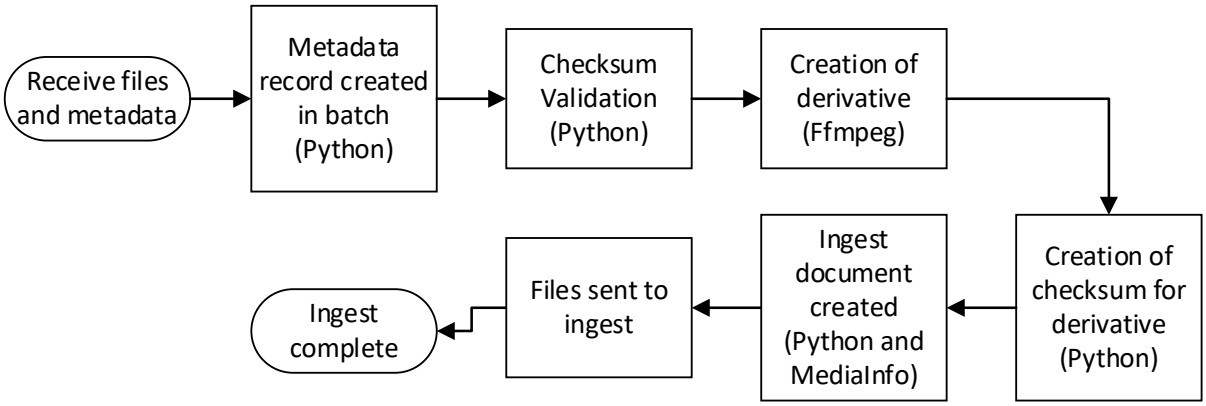


Figure 2: Hybrid manual and automated workflow

Checksum verification occurred at both the ingest package creation stage and ingest stage. The MySQL database [3] then stored the checksums, MAVIS ID, and a timestamp for when each step was completed.

The development and management of this workflow was distributed among staff members from a variety of functional and administrative areas. At the time, there was no dedicated staff member assigned to develop the workflow with integration into the NAVCC systems. The CPB-funded position was focused on overall project management and resolving any issues with the files as they were reported by the video lab supervisor.

In 2015, the Library hired a permanent AAPB digital project specialist responsible for the development and maintenance of the AAPB workflows. The digital project specialist quickly implemented changes to the initial workflow to accommodate new files being received through the AAPB project. This included shifting delivery from LTO-tape to hard drive, requiring pilot batches, adding a quality-control component, and requesting monthly file delivery rather than receiving all files upon a project's conclusion. The shift to hard drive from LTO was deemed necessary due to issues experienced with the LTO drive. Requiring pilot batches and running files through a basic QC profile in the Library's QC software, Baton [4], allowed both the Library and partner institutions to identify issues earlier in the process and relay these issues to the vendor in a timely manner. A cap of 25,000 files accepted per year was also implemented to prevent the accumulation of a backlog.

IV. Scalability, IT Modernization, and Challenges

As born-digital acquisitions increased, the Library hired two additional permanent digital project specialists in 2016 to process and ingest born-digital material outside the AAPB collections - one in the Moving Image Section, and one in the Recorded Sound Section. These two digital project specialists are devoted to born-digital collection work within their respective sections while also sharing and gaining insight from one another, with the goal of creating efficient processing workflows.

To meet this goal, the digital project specialists adapt to an ever-evolving technical infrastructure, modify workflows, and advocate for local technology updates based on observation, experience, and analysis of incoming collections. However, library-wide hardware and software changes are particularly challenging as the NAVCC workflows utilize different processes and systems than the rest of the Library. To tackle such constraints, the digital project specialists must often take the lead in resolving hardware and software changes and providing recommended solutions to improve and enhance current workflows.

Once the basic processing workflows (see Figure 2: *Hybrid manual and automated workflow)* were established for both the AAPB and general Moving Image collections, the staff at NAVCC began to investigate modernizing workflows for specific collections by implementing scalable, automated workflows. For an automated workflow to be successful, it should meet some minimum criteria: 1) consistently formatted machine readable metadata; and 2) consistent file naming and delivery.

Using a series of Python scripts and open source software such as mySQL, FFmpeg [5] and MediaInfo [6], automated workflows were developed based on individual processes found in the early processing workflows. Metadata records were also incorporated into the ingest packages for select collections if sufficient metadata was readily available. The next generation of this workflow evolution will include Baton quality control software [6] and a Dalet AmberFin transcoder [7] to replace FFmpeg during the creation of .mp4 files. The AmberFin is a shared resource at NAVCC with six servers and two transcoding engines per server, providing faster derivative and checksum creation for processing.

Each workflow generation brings unique challenges. Ultimately, a balance must be struck between multiple simultaneous workflows, the amount of available processing resources in the shared environment, and storage, and network bandwidth capabilities to write and move files.

At NAVCC, the evolution of digital moving image formats is ongoing – particularly related to files received from the entertainment industry – and we are prepared for these changes. The shift from SD to HD in television increased the file size by anywhere from 60% to 450% per hour, based on individual file characteristics, and we will see another significant increase in file size with the adoption of 8K resolutions. Increased file sizes create significant challenges in transferring files. For example, in 2017, most born-digital collections were sent via external hard drive to the Library. Currently, some collections are still received on a hard drive, some via SFTP, and still others are received via Amazon Web Services (AWS) or Aspera, a common entertainment-industry file transfer application. However, all these transfer mechanisms come with their own difficulties – hard drives require on-demand virus scanning and lengthy off-load time; SFTP transfers require IT department intervention; and Aspera transfers require navigating a rigorous security process that can take months to complete. Yet, despite these issues, the Moving Image Section is moving towards receiving more collections via SFTP and cloud transfer.

Processing storage and network bandwidth are also factors the Library must consider when attempting to increase our digital file transfer receipts. As noted previously, file sizes are increasing exponentially - a recent acquisition of a 4K motion picture was 781 GB for a 2-hour title - an average of 390 GB per hour. Conversely, processing storage is limited, relying on a constant movement of files during processing and ingest activities to remain viable. Downtime of systems and processing resources quickly result in an accumulation of files, consuming vital storage space. While the storage allocated for processing is generous - 80TB for AAPB and 100TB for other collections - large files require immense storage space. In this respect, we have more control as to what is in the processing space with hard drive transfers than with direct file transfers that are routed to specific directories for processing. If the processing space is full, we can opt not to offload an external hard drive, but direct file transfers will be received if there is available space. While some SFTP transfers occur overnight or in the early morning hours to minimize impact on overall network bandwidth, transfer systems such as Aspera utilize a “pull” wherein files are manually requested from an external source, often occurring during regular business hours, which are typically a peak network period.

Navigating current and changing infrastructure at the Library and NAVCC is a large component of the digital project specialists’ responsibilities. From advocating for infrastructure changes to navigating IT modernization and support issues, the digital project specialists work to maintain current workflows and optimize these workflows for future scalability.

V. The Current Born-Digital Moving Image Workflow Model

Over the past year, several Library-supported born-digital projects have seen an increase in file delivery and are planning to scale up even further. Scalability is critical for moving image processing to ensure maximized processing and leveraging of available resources to meet processing goals. The result is a flexible, scalable workflow model that can be adapted based on the characteristics of each collection (see Figure 3).

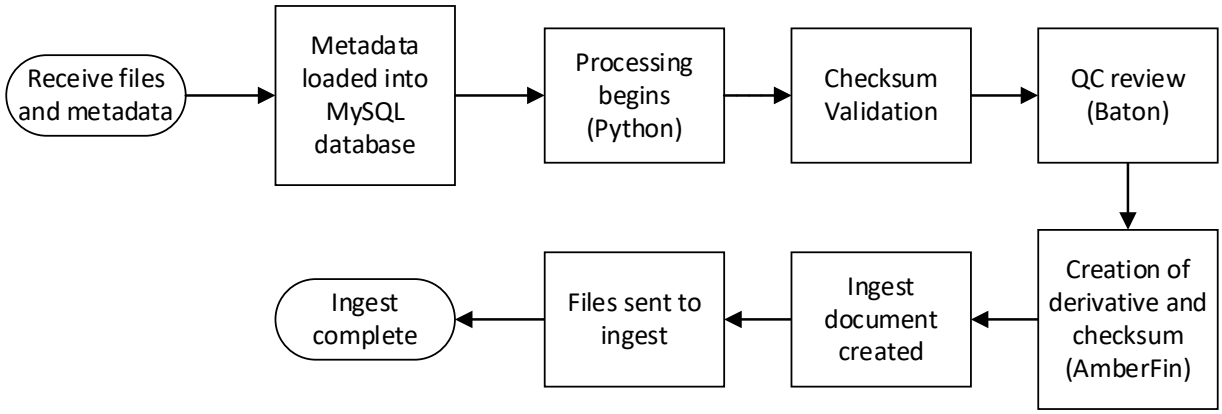


Figure 3: Current flexible and scalable born-digital moving image processing workflow model

Various tools, including Python scripting, MediaInfo, FFmpeg, OpenRefine and some limited direct file deposits, have been incorporated into AAPB workflows to prevent a backlog. This has resulted in two basic workflows for AAPB. The first workflow has manual elements to accommodate file deliveries without checksums and standardized file names. The second workflow is highly automated, and used for files delivered from a vendor, building on the flexible workflow model in Figure 3.

The Congressional Video project is another example of the desire to receive more content through sustainable scaling. The Library receives moving image files from the U.S. House of Representatives and U.S. Senate and is looking to expand content received, ensuring complete overlap with the National Archives and Records Administration. To do so, the Moving Image Section is working with the U.S. House of Representatives Recording Studio and U.S. Senate Recording Studio to standardize file delivery and file formats within the technological abilities and preferences of all project partners.

The U.S. Senate Recording Studio transitioned to solely file-based recordings in 2008. The Library began receiving daily file transfers in 2016, creating the impetus to develop the first automated processing workflow. Using a combination of Python, MediaInfo, FFmpeg, and a mySQL database as well as metadata provided by the Senate, this workflow validates file integrity via checksum verification, creates an .mp4 derivative file, issues the .mp4 checksum, gathers duration information, writes the metadata record, and generates the ingest file. This workflow has been in production since 2018 and is the foundation for automated workflows for other collections. The Moving Image Section is increasing the number of workflows that utilize these elements to enhance performance and allow the digital project specialists time to spend on projects that do not meet the requirements for automated workflows.

The extent of scalability for the non-AAPB collections is undergoing testing with the addition of the Congressional collections. Currently, each collection workflow uses its own virtual machine (VM), mostly due to the local transcoding function. However, with transcoding activities being moved to the Dalet AmberFin transcoder so processor-intensive work can be completed outside of the VM environment, the number of simultaneous processing workflows will increase.

This scalability is critical as the Moving Image Section looks to address the backlog in the born-digital moving image collections (non-AAPB), to create a sustainable, scalable processing model to ensure the best stewardship practice for the Library's collections and minimize or eliminate any processing backlogs.

VI. Conclusion

Performing analysis on the practical needs of born-digital projects and the impact of current policies that may not have born-digital workflows in mind are key to advocating the management of, if not more resources, then different approaches. Further, leveraging the expertise and experience of others working on similar projects creates a coalition when approaching management. Presenting a range of options increases the likelihood of finding a practical solution. Inevitably,

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

DOI: 10.1145/nnnnnnn.nnnnnnn

the volume of born-digital projects the Library is encountering will only increase, and establishing foundations now, in documentation and adapting workflows to changing circumstances, will surely ensure future success in this endeavor.

Preserving born-digital moving image content presents many challenges and opportunities. Ever increasing file sizes and storage requirements, technical infrastructure, and maximizing processing throughout are a few of these challenges. By implementing a scalable processing and preservation program with IT support, these challenges can be minimized if not mitigated entirely. Automated workflows and digital file transfers versus manual processing and hard drive transfers are two actions that have significant impact in increasing productivity while being mindful of storage and technological infrastructure limitation. Such interventions demonstrate the possibilities that can arise with thorough thoughtful planning and the inclusion of additional resources.

References

[1] OpenCube
[2] BagIt. https://www.ietf.org/rfc/rfc8493.txt
[3] MySQL database. https://www.mysql.com/
[4] Baton Quality Control Software by interra Systems. https://www.interrasystems.com/file-based-qc.php
[5] FFmpeg. https://ffmpeg.org/
[6] MediaInfo. https://mediaarea.net/en/MediaInfo
[7] Dalet Amberfin. https://www.dalet.com/products/amberfin/

TACIT PROCESSES

Qualitative Analysis Toward Bottom-Up Emulation Workflows

Eric Kaltman
California State University Channel Islands
USA
eric.kaltman@csuci.edu
0000-0002-7406-3827

Adam Larson
California State University Channel Islands
USA
adam.larson535@myci.csuci.edu

Abstract – This paper describes the use of a modification of qualitative grounded theory to analyze in-situ preservation workflows involving emulation techniques. The goal of this in-process work is to identify and delineate common tasks across the emulation of different classes of software objects through a unique approach based in bottom-up qualitative observation.

Keywords – Emulation, Digital Preservation, Qualitative Analysis, Grounded Theory

Conference Topics – From Theory to Practice; We're All in This Together.

I. INTRODUCTION

Software preservation workflows are becoming necessary within the greater orbit of digital preservation. Many legacy files, programs, and other born-digital materials in collections resist or would, in fact, be damaged by migration efforts. The use of virtualization methods, like emulation, to access, view, and manipulate legacy data in its original computing contexts is, therefore, necessary to preserve both the technical context of software's use and that of users' visual, tactile, and other embodied properties. Emulation, specifically, is becoming a common, catch-all term in digital preservation for any process that involves one computing context interpreting the data of another. While there is much work on emulation for software preservation, including many large, consortia helping to support emulation efforts, much of the discussion is not focused on how to proceed with emulation work but more on what that work, at a higher level, portends for the future of digital preservation. Working with virtualized environments to correctly configure and articulate legacy software dependencies and installations is (admittedly, according to many sources) an ad-hoc or bespoke affair. The technical nuances of different historical systems are highly varied and the network of dependencies for a given piece of software (and its dependent data) can grow daunting even for experienced users. Finding points of commonality across different classes of software, and different contexts of software study, would help to create a general set of procedures to build better workflows (and better-automated solutions) for emulation in preservationist contexts.

The purpose of this short paper is to lay out a methodology based in qualitative grounded theory for examining granular records of digital preservation activities involving emulated solutions and evaluating their common processes, including the mistakes and successes along the way. Although the use of emulation and virtualization is frequently advocated, it is rarely described (due to a lack of time or space) in enough detail for novices in the area to get started. The goal for this work is to take a closer look at the in-situ, tacit, and often overlooked processes that constitute digital preservation activities. The following sections provide some needed background on both emulation in preservation and qualitative methods. After that, the work proceeds with the organization of the initial study, explores early results, and then concludes with discussion and planned future work.

II. BACKGROUND

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

This section addresses, briefly, certain technical definitions that provide context to this work, the general desire in the community for these efforts, and notes on related emulation studies.

A. Emulation in Preservation

The use of emulation in libraries and other memory institutions has grown steadily since Rothenberg first posited the need for virtualization preservation solutions for born-digital software [1]. Generally, approaches to emulation make use of off-the-shelf (OTS) emulators or virtual machine managers (i.e. QEmu or Oracle's VirtualBox) that run on a host machine and allow the installation of guest operating systems or programs [2], [3]. The configuration and management of these OTS applications can become complex in many instances, with the practitioner needing experience with both guest and host OS installation procedures, networking configuration, data formatting and imaging, and general contemporary knowledge of the target data to be emulated [4]. As articulated by Acker, emulation in preservation work is conflated with general virtualization techniques to include any approach that allows one system to imitate the functionality of another [5]. Additionally, an emerging set of projects aims to make emulation workflows easier by abstracting the complex system configuration into the cloud. Systems like EaaSI and Olive allow expert practitioners to preconfigure environments on cloud-based servers and then view them through standard web-browsers [6], [7]. This study made use of both native OTS and cloud-based solutions.

B. The Need to Articulate Preservation Process

Although there is literature on the use of emulation in preservation, including in-depth analysis of emulation use case studies, emulation workflow design, and even qualitative studies of emulation workflows, there is also a consistent call within that same literature for better articulation of the requirements needed for emulation and software preservation activities [5, 8]. In many cases, institutions lack the technical capacity and staff necessary for comprehensive software preservation activities. Hagenmaier et al. explicitly call for more work on the finer details of software preservation workflows and the determination of commonalities across practitioner practice [8]. As noted above, many software preservation efforts are ad-hoc and institution-specific. The time and attention needed to disseminate explicit descriptions of highly varied workflows (each system has its own constraints and challenges) make most accounts that of individual trees instead of the forest. This work is positioned to begin the laborious process of recording, tabulating, and organizing disparate emulation use cases into a larger, generalized framework of practice that can inform future practitioners through the creation of training resources and computational support applications.

C. Related Work in Emulation

There are a few examples of emulation and software preservation workflows that inform this work. Acker investigated the workflows and management of the FCoP project, in which numerous GLAM institutions engaged with targeted emulation case studies. Acker used a modified grounded theory approach to qualify the larger domain of emulation practice [5], [9]. This present work seeks to look at similar processes but with a more granular focus. The goal is not to divine the larger categories of emulation use in GLAMs (Acker defined "preservation", "scholarly use" and "exhibition" as top-level concepts), but to model the day-to-day, minute-to-minute investigations and processes needed to recover to-be-emulated materials.

D. Grounded Theory and Diary Studies

The methodology used below is based on grounded theory (GT) with a data collection process akin to diary studies. GT is a qualitative analysis methodology that retrieves models and theories from raw data through a bottom-up, generative, and expandable process. The purpose of GT is to avoid a priori assumptions about a domain, and instead use observational data to derive concepts about it. There are many approaches to GT and this work most aligns with Corbin and Strauss due to their allowance for directed research questions and less restrictive methodology (for instance Glaser et al. prescribe specific analysis instruments that would not be applicable to this study's approach) [10]–[12]. In general, GT proceeds through distinct phases of initial conceptual coding, aggregating "selective" coding, and then theory "integration". Initial codes are derived from raw data and then compared and developed through "memoing", a process used to

elaborate on connections between concepts and their relationship to both the contexts of the described actions and their interrelationships. Another important aspect of GT is "theoretical sampling", in which insights from an initial data analysis identify further avenues for data sampling. This allows for the analysis to find new insights and then seek out new data to reinforce or contradict an emerging theory. The analysis ends with "saturation" when the researcher divines no new concepts or connections from the sampled data. Complimentarily, diary studies approaches collect longitudinal data from participants about a repeated set of activities through a self-reported diary [13]. In this study, the researchers recorded their daily efforts at software recovery through emulation.

II. Methodology

To generate the initial observation data for this project, three Computer Science undergraduate research associates (RAs) at California State University Channel Islands (CI) recorded their attempts to transfer and emulate software data from two sources: local materials stored on legacy media formats from the CI library, and a collection of interactive project backups donated by a well-known media arts program. The local data was completely unanalyzed, so its contents and requirements were determined during the study. The interactive arts projects had previously been studied in a different context related to file format profiles of game and entertainment development records [14].

The RAs had technical experience with emulators and virtual machine managers but not much experience with digital preservation workflows. This was a benefit in that many novice issues related to information gathering and configuration were cataloged. A potential negative is that some of their challenges might not occur in actual preservation practice, however, given that many institutions do not have well-developed digital preservation programs the RAs' technical backgrounds might be more developed than some library staff. Additionally, the veritable "clean slate" of the RAs' preservation knowledge caused them to find solutions and resources that had not occurred to the preservation expert that organized the study. Regardless, the recorded sessions do indicate numerous avenues for potential training topics and resources.

The RAs worked to recover any data they were interested in among the case set items. Specifically, RAs made use of a local EaaSI node, VirtualBox, and the MacOS SheepShaver and Basilisk II emulators [15], [16]. Observations were recorded daily for two months resulting in around 700 pages of notes. RAs were instructed to be as granular as possible and to identify all information sources consulted. The goal was to make target data objects available through emulation, however, there were no direct criteria for when an emulation task was considered complete.

After data collection, the notes were loaded into Altas.ti, a standard qualitative data analysis (QDA) tool [17]. The QDA allowed for simplified comparison between notes, automatic organization of codes, and aligning codes with analytic memos. Initial coding involved a reading pass through the notes followed by assigning conceptual and identifying codes to various quoted subsets. This will allow for future search and correlation analysis. Currently, the notes feature 1447 codes across 3588 quotations tied to 41 memos, however, the analysis is far from complete. The next step is to look through the assigned codes more deeply to find patterns and conceptual duplication. Many codes cover similar concepts (as noted below), and the goal is to arrive at a set of larger categories of preservation processes derived from codes that point toward conceptual unity across use cases.

III. From Practice To Theory: An Example With Host-Guest Data Sharing

This section will briefly detail the GT process as it is applied to the sharing of data between a host and a guest operating system. Typically, systems running in virtualization are sandboxed from the host environment. This means that data and file transfer into the guest environment needs to be mediated through some interface or connection between systems. While the analysis was not explicitly looking for this phenomenon, it arose from the initial coding with 5 related codes covering 75 quotations. These rough codes ("file transfer between host and guest", "host guest shared folder", "Guest Additions", "guest additions issue", and "VirtualBox guest additions") were then grouped under a "Host Guest Data Transfer" concept (seen in Figure 1). The "guest additions" refer specifically to a feature of the VirtualBox hypervisor that allows for modifications to be installed inside a virtual machine (VM) to

implement features that were not provided by the initial guest system. In this case, the additions allow for higher screen resolution than might have existed at the time, and for certain systems to access shared memory locations to enable shared folder access.

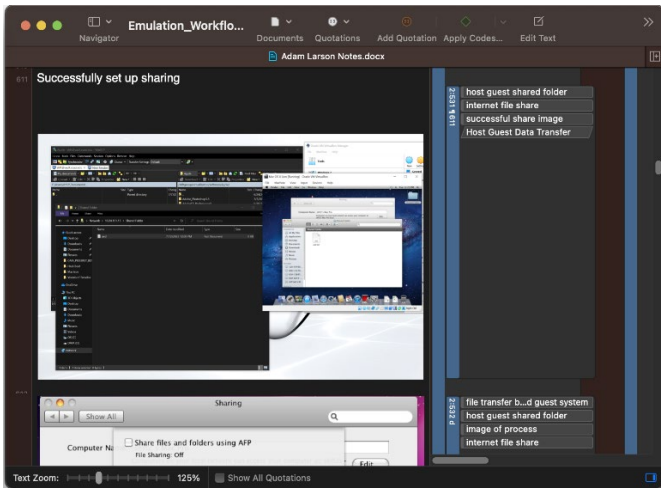


Figure 1. Altas.ti interface with codes applied to screenshot of successful data transfer

In looking at the details of the coded quotations, it is possible to cross-reference these conceptual codes with identifying codes that describe the operating system and tools being used. In this case, the codes correlated with the use of VirtualBox to virtualize Microsoft Windows (specifically ME, XP, Vista, Server 2008, 7, and 10) and MacOS X (Lion, Snow Leopard) environments, along with Sheepshaver and Basilisk II (for System 7, MacOS 8 and 9). To proceed further, GT methods then inquire into the specific dimensions and properties of observed actions and interactions. These are then placed in a larger context to hopefully intuit some emerging theory of process. One potential property of the "Host Guest Data Transfer" was the specific interface needed to allow it. Based on the coded quotations, there appeared to be four primary methods of importing data into a guest environment:

1. Shared folders that allowed for a storage location on the host to be mounted inside the guest.
2. Shared network folders connecting the host and guest machines through a virtual network controller.
3. Allowing the guest to connect to the Internet and remotely download files.
4. Loading the data into a disk image and mounting it in a virtualized media drive.

While these approaches were not decided in advance, they emerged as a result of the interaction between the practical needs of the RAs for recovering specific objects and the available features provided by the virtualization technologies. The use of GT allowed for the organic detection of specific patterns of preservation actions and interactions that corresponded with the larger "process" of host to guest data transfer. Here, the analysis highlighted data transfer as an area of contention among the RAs (in that they repeatedly noted difficulties with consistent data sharing) and what the general solutions appeared to be, given the case items.

Further, it is possible to view the quotations linked to these methods and divine potential dimensions of the data transfer concept, like the symmetry of the methods used. In the case of methods 1 and 2, there was a symmetric link established between the host and guest that allowed for transfer into and out of the guest environment. However, methods 3 and 4 are unidirectional, in that they allow for data to go into the guest environment without a complementary retrieval mechanism. In fact, method 4 was the primary means used by the EaaSI system for inserting data into environments highlighting a potential difficulty with cloud-based emulation solutions vis-à-vis locally executed ones. This dimension of "symmetry" in data transfer processes is then a potential new site of analysis as the concept can be compared with the literature for further elaboration and validation.

Additionally, it is also possible to look at the knowledge context within which these data transfer methods are embedded. Since the RAs also recorded where they researched the data sharing methods, a network of online and textual documentation, individual experimentation, and online tutorials and videos prefigures the combined knowledge necessary to engage, as a preservation practitioner, with the "Host Guest Data Transfer" concept. Continued work on related preservation tasks would likely position this concept relative to other processes needed for the emulation of target data objects. From this, a general theory of emulated preservation techniques could then emerge.

Finally, there is serendipity and surprise in the malleability of the GT approach that finds meaning in "mundane" minutia. When working through sharing method 1, one RA realized that they needed to move files from the guest shared folder into a local one to

avoid permission and access issues. Another RA discovered that method 2 necessitated removing significant network security features from the *host system* for guest access to be possible. These new notions related to "permission" and "security" might now be potential vectors for dimensional analysis.

IV. Promise, Limitations And Future Work

The preceding example highlights how close attention to practitioner activities can reveal deeper relationships between seemingly disparate preservation targets and points to the potential for subject agnostic knowledge sharing. In the example above, the RAs were working with data from disparate sources but they all still needed to find some way to get that data, once acquired, into the emulated environment. The GT approach required that the raw details of the process be reconsidered in comparative and generalized contexts, and it was through this consideration that patterns started to emerge. However, this work is developing as there are hundreds of codes to aggregate and process.

Contrarily, some limitations of the current study must be noted, including items that will change for future studies. The sample size, while extensive in activities, was limited in participants. The RAs worked for a combined 960 hours, and the resulting notes are rich in specific details relating to a variety of emulated environments. However, since the RAs were students and not trained preservationists, it is unclear if some of the specific procedures or issues encountered might simply not occur with more experienced practitioners. A caveat here is that GT methods are designed to address sampling issues by allowing for "theoretical sampling" based on progressive findings. It would be feasible to add the subject position of the individual as a dimension of the analysis and compare practitioner experience with the execution of preservation tasks. Additionally, embedding more self-reflection into the process would be beneficial. The researchers did not proceed with GT analysis until after the initial data collection interval ended. It would have helped the study to begin coding and analysis during data collection to steer the RAs toward fruitful pathways. The next steps for this research are to proceed with constructing a model of both the dependencies and related processes incumbent on the emulation of software and software-dependent data objects. Current progress is promising and there are likely to be more unlikely commonalities discovered across the documented use cases, effectively creating theory from practice.

V. Acknowledgements

We wish to thank the additional RAs, Desireé Caldera and Morgan McMurray, and the CI Summer Undergraduate Research Fellowship (SURF) for providing support to this work.

VI. References

[1] J. Rothenberg, "An experiment in using emulation to preserve digital publications," 2000, Accessed: Sep. 16, 2016. [Online]. Available: http://www.kb.nl/sites/default/files/docs/emulationpreservationreport.pdf

[2] "Oracle VM VirtualBox." https://www.virtualbox.org/ (accessed Sep. 19, 2021).

[3] "QEMU." https://www.qemu.org/ (accessed Sep. 19, 2021).

[4] D. S. Rosenthal, "Emulation & Virtualization as Preservation Strategies," 2015.

[5] A. Acker, "Emulation practices for software preservation in libraries, archives, and museums," *J Assoc Inf Sci Technol*, p. asi.24482, May 2021, doi: 10.1002/asi.24482.

[6] "EaaSI," *Emulation-as-a-Service Infrastructure*. https://www.softwarepreservationnetwork.org/emulation-as-a-service-infrastructure/

[7] M. Satyanarayanan *et al.*, "Olive: Sustaining executable content over decades." XSEDE, 2014.

[8] W. Hagenmaier, C. Williford, L. Work, J. G. Benner, S. Erickson, and M. Lassere, "Supporting Software Preservation Services in Research and Memory Organizations," Software Preservation Network, White Paper, 2022.

[9] A. Acker, "Accessing Software: Emulation in Information Institutions." Rochester, NY, Apr. 06, 2023. Accessed: Jun. 08, 2023. [Online]. Available: https://papers.ssrn.com/abstract=4450195

[10] K. Charmaz, *Constructing Grounded Theory*, 2nd Edition. London ; Thousand Oaks, Calif: SAGE Publications Ltd, 2014.

[11] B. G. Glaser, A. L. Strauss, and E. Strutzel, "The discovery of grounded theory; strategies for qualitative research," *Nursing research*, vol. 17, no. 4, p. 364, 1968.

[12] J. Corbin and A. Strauss, *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications, 2014.

[13] K. Salazar, "Diary Studies: Understanding Long-Term User Behavior and Experiences," *Diary Studies: Understanding Long-Term User Behavior and Experiences*, Jun. 05, 2016. https://www.nngroup.com/articles/diary-studies/ (accessed Nov. 30, 2022).

[14] E. Kaltman, R. Lorelli, A. Larson, and E. Wolfe, "Organizing a Content Profile for a Large, Heterogeneous Collection of Interactive Projects," in *2021 IEEE International Conference on Big Data (Big Data)*, Dec. 2021, pp. 2231–2239. doi: 10.1109/BigData52589.2021.9671904.

[15] C. Bauer, "SheepShaver," *SheepShaver: An Open Source PowerMac Emulator*. https://sheepshaver.cebix.net/ (accessed Mar. 10, 2023).

[16] C. Bauer, “Basilisk II,” *Basilisk II: An Open Source 68k Macintosh Emulator*. https://basilisk.cebix.net/ (accessed Mar. 10, 2023).

[17] “ATLAS.ti,” *ATLAS.ti*. https://atlasti.com (accessed Mar. 10, 2023).

POLICIES, RISKS AND STRATEGIES:

A File Format Debate

Sam Alloing
National Library of the Netherlands (KBNL)
sam.alloing@kb.nl
0000-0002-1254-1483

Valentijn Gilissen
Data Archiving and Networked Services (DANS)
the Netherlands
valentijn.gilissen@dans.knaw.nl
0000-0003-2399-7598

Leslie Johnston
National Archives and Records Administration (NARA)
United States
leslie.johnston@nara.gov
0000-0001-9908-0183

Kate Murray
Library of Congress (LoC)
United States
kmur@loc.gov
0000-0003-1325-0829

Tyler Thorsted
Brigham Young University (BYU)
United States
thorsted@byu.edu
0000-0003-0292-0962

Paul Wheatley
Digital Preservation Coalition (DPC)
United Kingdom
paul@dpconline.org
0000-0002-3839-3298

The digital preservation community has been developing approaches to preserving the meaning of digital content for a number of decades. But questions still remain as to the most accurate, practical, timely and cost effective way of keeping our data usable. Collating and presenting file format policies from several organizations triggered a lively panel discussion in early 2023. This panel session will build on the success and popularity of that debate by bringing in new voices and topics raised by the audience. This subject is a critical one to understand if we are to be successful in preserving our data for future generations.

Keywords – File formats, file format policy, file format assessment, preservation planning, preservation strategy

Conference Topics – Sustainability: Real and Imagined, From Theory to Practice

1. BACKGROUND

The International Comparison of Recommended File Formats [1] collates file format policies from 28 organizations from around the world. Paul Wheatley published a blog post which questioned a number of aspects of this work [2]. On February 9th 2023 Sam Alloing (KBNL) moderated a panel debate between Valentijn Gilissen (DANS) and Paul Wheatley (DPC), entitled "Do unacceptable file formats exist?". The event drew a crowd of 200 people and provoked an almost overwhelming degree of comment and engagement from the audience. This panel session aims to build on the success of that debate by bringing an extended panel of diverse opinions to the iPres Conference.

2. CONTRASTING APPROACHES TO PRESERVATION STRATEGY

A strategy used for file format preservation is the use of preferred file formats. In this strategy a file format is identified as preferred if it complies with some defined criteria. DANS has such a Preferred Formats policy [3], as does NARA [4], the LoC [5], KBNL [6] and others. For both DANS and NARA, a file acquired in a non-preferred file format is migrated to a preferred file format if possible and the original retained. At the KBNL, all file formats of a publisher are allowed and preserved. The KBNL's policy is to assign file formats a 'knowledge level'. This is the status of a file format in the repository and indicates what preservation operations are possible. For example the first level is 'stored file', this means that the file is only bit-preserved. The third and last level is known 'file format', where the results of identification, validation and technical metadata extraction can be interpreted and guidelines for each format have been formulated. Preferred formats are typically identified through generic criteria such as

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

age, tool support, complexity, documentation, risk and context. Examples include the LoC Recommended Formats Statement evaluation matrix and the NARA Risk Assessment Matrix [5, 7]. The resulting data would then be used to determine file format policy and ultimately which formats to migrate. An opposing view was offered by van der Knijff who argued that such risk factors were largely theoretical [8]. Rosenthal argued that "format obsolescence is a rare problem" due in large part to the availability of open source rendering tools [9]. Just Solve has focused on documenting and web archiving sources of information on file formats [10].

3. Broadening The Debate

The variety of perceptions in the world of digital preservation may seem to conflict with each other. Having an open debate about these subjects provides a fruitful basis for sharing knowledge and gaining consensus. This panel session will continue, broaden and extend the debate held in February 2023. It will incorporate the diverse viewpoints of several members of the audience of that original debate. Leslie Johnston (NARA) brings the stark challenges of the long-term preservation of an ominously large range of different file formats. Kate Murray (LoC) brings experience of researching and accessing file formats through her leadership of the Sustainability of Digital Formats and Recommended Formats Statement. Tyler Thorsted (BYU) brings a track record of contributions to the PRONOM and Just Solve registries. Leslie, Kate and Tyler will join the panelists of the original debate: Valentijn Gilissen oversees the file format guidelines of the Dutch national centre of expertise and repository for research data (DANS) in his role as preservation officer. Paul Wheatley is Head of Research and Practice at the Digital Preservation Coalition. Sam Alloing (KBNL) actively contributed to the Guide to Preferred File Formats of the Dutch Digital Heritage Network (DDHN) and the analysis of the File Format Lifecycle also from the DDHN.

4. Format Of The Panel

Following short introductions from each of the panelists the session will move to a question and discussion format, moderated by Sam Alloing. It will focus primarily on questions of file format policy and digital preservation strategy. The considerable text chat from the February panel discussion will be used as a source of topics for discussion. The panel will ensure strong audience participation by both accepting questions from them and posing live poll questions to them. This will provide an impression of the state of play for preservationists represented at iPres alongside the viewpoints of the panel members. Activating the audience with poll questions demonstrated a meaningful and active discussion in the February debate, so we would like to replicate that approach here. The format allows remote and in-person participation.

Key questions for discussion by the panel members include: 1) What are the criteria for file format assessment in the global and institutional contexts? 2) Is there such a thing as a "good, bad or unacceptable" format? 3) What goes into risk assessment for file formats? 4) How do file format risks compare to other risks in the field of digital preservation? 5) What strategies are used for assessing file format risks?

5. REFERENCES

[1] International Comparison of Recommended File Formats [Online] https://openpreservation.org/resources/member-groups/international-comparison-of-recommended-file-formats/

[2] Wheatley, P. "File format recommendations..." Blog post, DPC, [Online] https://www.dpconline.org/blog/file-format-recommendations

[3] Laagland et al, "White paper on preferred formats" http://doi.org/10.5281/zenodo.4518486, p 19

[4] DANS Preferred Formats guidelines [Online], https://dans.knaw.nl/en/file-formats/

[5] NARA Format Guidance for the Transfer of Permanent Electronic Records [Online], https://www.archives.gov/records-mgmt/bulletins/2014/2014-04.html

[6] Library of Congress Recommended Formats Statement [Online], https://www.loc.gov/preservation/resources/rfs/

[7] NARA Digital Preservation Framework [Online], https://github.com/usnationalarchives/digital-preservation

[8] Van der Knijff, J. "Assessing file format risks: searching for Bigfoot?", Blog post, Bitsgalore [Online] https://www.bitsgalore.org/2013/09/30/assessing-file-format-risks-searching-bigfoot

[9] Rosenthal, D. "Format Obsolescence: Assessing the..." https://web.stanford.edu/group/lockss/resources/2010-06_Format_Obsolescence.pdf

[10] ArchiveTeam File Format Wiki [Online] http://fileformats.archiveteam.org/wiki/Main_Page

Approaches to Digital Preservation Product and Service Sustainability

Comparing alternate approaches

Jack O'Sullivan
Preservica, UK
Jack.osullivan@preservica.com0000-0002-0306-761X

Oya Y. Rieger
ITHAKA, USA
oya.Rieger@ithaka.org

Kelly Stewart
Artefactual, Canada
kstewart@artefactual.com

Thib Guicherd-Callin
LOCKSS, USA
thib@cs.stanford.edu
0000-0002-6425-4072

David Giaretta
Primary Trustworthy Digital Repository Authorisation Body Ltd, UK david@giaretta.org

William Kilbride
Digital Preservation Coalition
william.kilbride@dpconline.org

Abstract – How do we apply the lessons of ongoing evaluations of digital preservation sustainability within single institutions to the products and services on which this sector increasingly depends? The speakers will look at this key question from different viewpoints to pool best practice and explore the issues to ensure the community can expect more durable systems however they are delivered.

Keywords – Sustainability, Products, Services, Standards

Conference Topics – Sustainability: Real and Imagined

1. Introduction

Sustainability is of course critical in the field of Digital Preservation and has been the subject of many surveys, reports, evaluations, standards and papers, almost always applying to individual preservation programs. However, as the field has evolved, almost all preservation initiatives will select from the products available and implement one of these, placing a burden on the providers of these systems to ensure that what they provide delivers the sustainability demanded.

The sad demise of the DPN system in 2018 [1] demonstrated that even well-funded, widely applied systems are not immune from failure and such events reflect badly on the whole sector.

This panel will discuss the sustainability of these products and services, where sustainability is considered in its widest form. It will take input from academic studies, product suppliers and standards experts to examine the approaches adopted by these providers, comparing models and setting expectations for the whole sector. The speakers have been selected to represent different approaches to achieving the same endpoint – digital preservation products and services that sustain for the long term.

2. Previous Sustainability Evaluations

Many approaches exist to evaluate the quality of specific digital preservation programs. Most of these center around standards as summarized by the Digital Preservation Coalition [2]. These cover many alternatives, from heavyweight specific and general ISO standards, sector specific standards, standards on part of the challenge like metadata or storage, to lighter weight evaluations such as the DPC Rapid Assessment Model, NDSA Levels of Preservation, and CoreTrustSeal.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

These all have their place, but they are mainly aimed at organizational program sustainability rather than that of a product or service that will be offered to tens, hundreds, or thousands of organizations. They also don't cover supplier longevity issues like financial stability, internal knowledge management, or environmental impact. Most importantly they are not appropriate for the evolving market of new users who just want to have confidence that a system works and do not have the funds or knowledge to do a full audit.

3. Product and service Sustainability Factors

The sustainability of a digital preservation product or service should be judged against the following criteria:

- Data. How is the content sustained?
- Software. How is the software used kept in production and up to date even after the supplier ceases to exist?
- Operations. How do the processes required to operate the product or service continue?
- Knowledge. How does the organization avoid relying on a few key individuals?
- Financial. What is the commercial model to ensure the long-term viability of the services?
- Governance. How does the organizational structure ensure good practice is maintained?
- Environmental. Is the organization demonstrating their responsibility to ensure a minimum environmental footprint?

These questions apply to all organizations active in this field, whether the source code is open, closed or escrow, whether the ownership is commercial, community or academic, and whether the solution is delivered as installable software or a hosted service.

4. PANEL SPEAKERS

Oya Rieger (Senior Strategist, ITHAKA)

Oya has authored several reports on the state of the Digital Preservation landscape [3] and will use the findings of her recent paper "The Effectiveness and Durability of Digital Preservation and Curation Systems" [4] to show how organizations with very different commercial models deliver sustainability with the digital preservation context. Oya also contributed to the report on the failed DPN initiative.

Thib Guicherd-Callin (Program Manager, LOCKSS)

Thib will present on the sustainability of the community of services built around LOCKSS, the open-source software, which has served distributed digital preservation to scholars, libraries, memory organizations, and publishers for two decades.

Jack O'Sullivan (Innovation Engineering Lead, Preservica)

Jack will present on how Preservica has delivered sustainable digital preservation technologies for 20 years, drawing on their recently published Sustainability Charter [5].

Kelly Stewart (Chief Archivist, Artefactual)

Kelly will present on Artefactual's approach to sustainability as a company providing commercial services around the open-source software it stewards.

David Giaretta (Director and Lead Auditor, Primary Trustworthy Digital Repository Authorisation Body Ltd)

As one of the leaders in Digital Preservation international standards, David will draw on ISO 16363 (Audit and certification of trustworthy digital repositories) to highlight those elements that should apply to product and service providers.

William Kilbride (Executive Director of the Digital Preservation Coalition (DPC))

William will moderate the session, reflecting his role making sure that the DPC members can rely on all products and services across this emerging sector.

5. REFERENCES

[1] Why Is the Digital Preservation Network Disbanding?, Roger Schonfeld, https://scholarlykitchen.sspnet.org/2018/12/13/digital-preservation-network-disband/

[2] Digital Preservation Handbook: Standards and best practice, Digital Preservation Coalition, https://www.dpconline.org/handbook/institutional-strategies/standards-and-best-practice

[3] The State of Digital Preservation in 2018, Oya Y Rieger, https://sr.ithaka.org/publications/the-state-of-digital-preservation-in-2018/

[4] The Effectiveness and Durability of Digital Preservation and Curation Systems. Oya Y. Rieger, Roger C. Schonfeld, Liam Sweeney, https://sr.ithaka.org/publications/the-effectiveness-and-durability-of-digital-preservation-and-curation-systems/

[5] A Charter for Long-term Digital Preservation Sustainability, Preservica, https://cdn2.assets-servd.host/preservica-core/production/resources/A-Charter-for-Long-term-Digital-Preservation-Sustainability.pdf

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

CREATING DIGITAL PRESERVATION PLANS

Leveraging Expertise Across Your Organization

Jeanne Kramer-Smyth
World Bank Group
USA
jkramersmyth@worldbankgroup.org
0000-0002-5689-8409

Thomas Gkremo
World Bank Group
USA
tgkremo@worldbankgroup.org
0009-0006-6717-2298

Sherrine Thompson
World Bank Group
USA
sthompson3@worldbankgroup.org
0009-0007-3668-4511

Abstract – The creation of Digital Preservation Plans requires leveraging a wide range of archival expertise. Our panel will discuss each of the components we have identified for inclusion in our preservation plans, along with specific skills and knowledge we depend upon from different parts of our team. Learn how we use a standard framework and leverage the expertise and enthusiasm of our Appraisal, Transfer, Ingest, and Arrangement and Description teams to create thorough and functional Digital Preservation Plans. Session will include recommendations of how to apply our approach at your institution.

Keywords – digital preservation, leveraging expertise, digital preservation planning

Conference Topics – From Theory to Practice

I. INTRODUCTION

The World Bank Group Archives (WBGA) has developed an approach for creating Digital Preservation Plans for each record type slated to be ingested into the Digital Vault (the WBGA's digital preservation platform). We define a record type as the intersection of a digital format and source business unit. For each record type, we want to ensure that we have done our due diligence to define and document the processes that will guide us from identifying records for preservation through to long term access to those records.

Building on the iPRES 2019 panel in which two WBGA's staff participated (The People and Processes of Digital Preservation), this panel will discuss how we have transitioned from the design phase to the implementation phase for the Digital Vault.

Much of the hands-on work of digital preservation takes place outside of technical platforms. It requires methodical coordination and a deep understanding of each record type we need to preserve. The Digital Preservation Plans discussed by this panel seek to both distribute the work necessary to preserve born-digital permanent records, but also to acknowledge that we need all the branches of archival expertise in our organization to be successful. We will discuss methods used to gain buy-in from our broader team and how having a formalized structure for contributions helps us in our ongoing work to ingest and preserve key digital records of the World Bank.

A. Digital Preservation Plan Components

Creation of a Digital Preservation Plan requires the following components:

- Sample Data: a set of representative sample data.
- Appraisal and Selection Criteria: While the WBGA depends on our record schedules to identify records for long term preservation, digital records often require additional criteria be applied during the selection process.
- Metadata Profile: list of attributes that we would like to assign at the digital object level in Digital Vault.
- Content Manager Digital Transfer Values: Values needed to create a Digital Transfer in Content Manager (our union catalog of both analog and digital records in custody of the WBGA).

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

- Transfer Technical Design: How to transfer records to a WBGA controlled staging area
- Ingest Technical Design: Any special requirements for ingesting records into Digital Vault.
- Digital Vault Destination Folder: Where should records be placed in the Digital Vault hierarchy?
- Arrangement and Description Unique Guidelines: guidelines unique to this record type that will support arrangement and description, often an extended time after the original ingest.
- Format Preservation Research: Preservation challenges related to the format of files associated with this record type, along with recommended action plans to ensure long-term access.

The first portion of the panel will focus on defining each of the components listed above and how we came to determine that each component was a necessary part of a Digital Preservation Plan.

B. Drill Down into Details

In the second section of our panel presentation, we will deep dive into selected examples of a few of the more complex components, such as:

- Appraisal and Selection Criteria
- Metadata Profile
- Arrangement and Description Unique Guidelines

This will give our panelists the opportunity to highlight a success story of how each of these components demonstrated their value in our digital preservation program.

We will also review the final product of all the Digital Preservation Plan work for a single record type: a "Digital Preservation Action Plan" which combines all the decisions into a single reference document for that record type to be used by staff across the WBGA team.

We will also discuss:

- An overview of the WBGA team configuration
- Tips on getting buy-in from our team
- Examples of each component
- The evolution of this living process to create these plans. We are learning as we work and still have many plans yet to be created.
- Suggestions on how this approach might be implemented at other organizations

C. Q&A

Part of the panel time will be reserved for discussion and answering questions from the audience.

FROM THEORY TO PRACTICE: UNDERSTANDING THE EVOLUTION OF A DIGITAL PRESERVATION PROJECT FROM CONCEPTION TO FINAL REPOSITORY

Some real cases

Almudena Caballos
Universidad Complutense de Madrid
Spain
acaballo@ucm.es

Chris Knowles
Churchill Archives Centre,
University of Cambridge
UK
Chris.Knowles@chu.cam.ac.uk

Kate Cawthorn
University of Calgary
Canada
kathryn.cawthorn@ucalgary.ca
ORCID 0000-0003-0271-400X

Antonio Guillermo Martinez
LIBNOVA SL
Spain
a.guillermo@libnova.com

Maria Fuertes
LIBNOVA SL
Spain
mfuertes@libnova.com

Abstract – Usually, the resulting digital preservation project is very different from its conception.

The theory defined before starting a digital preservation project serves as a guide for the beginning of the project, but must be flexible enough to be adapted throughout the implementation to fit the real needs of the organization.

In this panel, representatives of institutions from different GLAM-UR sectors from different countries will speak from their own experience about the evolution of a digital preservation project from its theoretical conception to its real practical implementation.

Keywords – Digital Preservation Project, Digital Repositories, Implementation

Conference Topics – From Theory to Practice, We're All in this Together

A Digital Preservation Policy [1] is the mandate for an archive to support the preservation of digital records through a structured and managed digital preservation strategy. The policy details why selected material needs to be preserved; the strategy defines how this will be implemented. But this strategy must be flexible enough to allow organizations to tailor their decisions to their needs during the implementation process.

Planning and implementing a digital preservation project requires consideration of many aspects. There are some manuals such as the DPC Digital Preservation Handbook [2] or the CHIN Digital Preservation Toolkit [3] that provide practical and internationally authoritative guidance to help organizations assess their own digital preservation needs and guide them in developing digital preservation policies, plans and procedures. However, it is important to understand that a digital preservation project is not a closed endeavor, but will evolve throughout the process.

In this panel, professionals responsible for digital preservation from the following institutions will exchange their experiences implementing a preservation project in different kinds of libraries and archives: *Universidad Complutense de Madrid,* Spain; *The Churchill Archives Center at the University of Cambridge,* UK; and *University of Calgary,* Canada. They will share the evolution of a digital preservation project from its theoretical conception to its real

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

practical implementation, their workflows, and some useful insights for anyone in the same situation.

LIBNOVA is the common denominator among the different organizations, and its role will be to serve only as a moderator of the panel session.

1. Panel Discussion Topics

The panel will discuss the following topics and questions:

- Key aspects to consider when planning a digital preservation project.
- Organization of the preservation team, roles, and coordination between the different areas involved.
- Process and methodology adopted to select the material to be preserved.
- Workflows and how they have been defined.
- General overview of how the project has evolved from its initial conception to the present time.

2. REFERENCES

[1] Developing a Digital Preservation Policy. The National Archives. https://cdn.nationalarchives.gov.uk/documents/digital-preservation-policies-cons-draft-0.9.pdf

[2] Digital Preservation Handbook, 2nd Edition, https://www.dpconline.org/handbook, Digital Preservation Coalition © 2015.

[3] Canadian Heritage Information Network Digital Preservation Toolkit https://www.canada.ca/en/heritage-information-network/services/digital-preservation/toolkit.html

The Curricular Asset Warehouse At The University Of Illinois

A Digital Archive's Sustainability Case Study

Karin Hodgin Jones
University of Illinois at Urbana-Champaign
USA
khodgin2@illinois.edu

Dr Jimi Jones
University of Illinois at Urbana-Champaign
USA
jjones7@illinois.edu

Robyn Bianconi
University of Illinois at Urbana-Champaign
USA
biancon1@illinois.edu

Liam Moran
University of Illinois at Urbana-Champaign
USA
moran@illinois.edu

Abstract – What happens to the devices that host digital objects – hard drives, monitors, computer peripherals, storage media – when it is time to upgrade digital preservation environments and workflows? Each step of the production and stewardship of digital objects requires devices and software that have short life cycles and multiple drivers of ever faster obsolescence. These devices flow out of digital repositories and contribute to the fastest growing waste stream of the 21st century: electronic waste or "e-waste."

The Center for Innovation in Teaching and Learning (CITL) at the University of Illinois at Urbana-Champaign is currently working with the head of the university's Sustainable Design program to perform a case study of the sustainable management of its large volumes of digital video and image content production and preservation, within an analysis of its institutional purchasing and waste management paradigms. The purpose of this analysis is to determine how device obsolescence at CITL can be mitigated to avoid future costs and to minimize the department's contribution to the global e-waste problem.

Keywords – Media Asset Management; Sustainability; Electronic Waste

Conference Topics – Sustainability: Real and Imagined; From Theory to Practice

I. Introduction

Over the past five years, the Center for Innovation in Teaching and Learning (CITL) at the University of Illinois at Urbana-Champaign has been developing a system called the Curricular Asset Warehouse (CAW), which is a suite of software that serves as the backbone of its production and archival needs. CAW uses several open-source software tools to serve as an all-in-one production, cataloging, preservation and discovery tool. CAW is useful to CITL's media producers and archivists because it helps facilitate collaboration on media production projects while also minimizing extraneous data in CITL's digital storage.

Digital audiovisual files are large and resource-intensive to manipulate and store. Because CAW integrates software and hardware to maximize the efficiency of its audiovisual production and storage, CITL is participating in a case study to determine how environmentally friendly the CAW software is. The sustainability case study also assesses CITL's media production and preservation workflows as well as the hardware the department uses for these purposes.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

This panel lays out the development of CAW and describes the current case study of CITL's incidental and intentional sustainability practices. The study began with a general analysis of the three classic pillars of sustainability: environmental impact, social equity, and economic benefit that preserve the potential of sustained economic, environmental and social benefit into the future. The primary area of inquiry was e-waste impacts related to hardware and software choices. Many of these choices for procurement and responsible stewardship of electronic devices at CITL were rooted less in a conscious selection for lowest environmental impact but instead were driven by access to reusable or repurposable, high-quality electronics and the ability to maintain uniformity across team access. The case study depicts the choices made by the CAW development team, within specific budget constraints, as an accretive process over time, within a state institution. The findings are a start at analyzing many of the current methods of e-waste management, how and why organizations make the choices they do for device procurement, reuse and discard, and where there can be greater flexibility of choice toward more sustainable outcomes.

The case study analysis focuses on hardware, core devices and peripherals, and software used by CAW between the (hot) production stage, in the accessible distribution and archive stages and, through the long-term (cold) storage process. The environmental impacts are determined by the length of time electronics are used before they need to be replaced and the energy efficiency of devices and electronic resources. The equity part cannot be overlooked though it is a fixed feature; everything they do is open source. The team is committed to making their methods and documentation of equipment accessible and usable by people across multiple organization types, within primary and higher education spaces and other organizations. This research grows from that spirit of collaboration and open access.

II. About The panel

This panel will feature professionals from the University of Illinois at Urbana-Champaign, who will discuss the digital media asset management practices at the Center for Innovation in Teaching and Learning (CITL), a high-throughput video production unit, and how those practices relate to environmental sustainability.

Robyn Bianconi will talk about the history of asset production and management at CITL, from the days of mini-DV tape video capture to the current era of tapeless production and LTO tape storage. Robyn will give context for the development of CAW.

Jimi Jones and Liam Moran will talk about how CAW's role in the digitization and preservation workflows at CITL are an effective strategy for minimizing CITL's digital storage needs, and, by extension, the amount of electronic waste produced by the department. Jimi will also discuss CITL's choice of LTO for digital preservation, its utility as air-gapped storage that needs little maintenance and how it helps to save space (and write-cycles and longevity) on CITL's production servers.

Karin Hodgin Jones will talk about CAW as a case study in sustainable management of large volumes of digital video and image content production and preservation, within an institutional purchasing and waste management paradigm with foresight into the drivers of device obsolescence to mitigate future costs and redundancies.

While this panel is in dialogue with current waste study and standards development theory, the panelists will also give participants real-world sustainability tips and solutions that they can implement. These tips and suggestions will be informed by the design of the CAW hardware and software suite and can be a roadmap for how to locate reuse strategies at multiple scales within and between institutions.

From Community-Supported Values to Action

Operationalizing the Digital Preservation Declaration of Shared Values

Hannah Wang
Educopia Institute
USA
hannah.wang@educopia.org
0000-0002-6676-1254

Jess Farrell
Educopia Institute
United States
jess.farrell@educopia.org
0000-0001-9794-6908

Courtney Mumma
Texas Digital Library
USA
c.mumma@austin.utexas.edu
0000-0003-1394-5006

Sibyl Schaefer
University of California, San Diego
USA
sschaefer@ucsd.edu
0000-0002-7292-9287

Abstract – Operating from a foundation of shared values, the community-supported digital preservation services represented in the Digital Preservation Services Collaborative (DPSC) empower stewards of digital content to make informed decisions by offering transparency and accountability. These values have become increasingly important as resources for digital preservation fail to meet the needs of organizations, forcing many mission-critical digital preservation activities to be outsourced to commercial providers. In the DPSC Planning Project, this group of mission-aligned service providers are working to establish closer and more intentional collaboration between their organizations, in order to guarantee the continued availability of services that prioritize transparency and accountability to the cultural heritage organizations they serve. In this interactive panel, partners from the DPSC Planning Project will discuss the importance of the shared values for digital preservation good practice, how they have enacted these values within their organizations, and other project findings.

Keywords – Digital preservation; collaboration; values; transparency; accountability

Conference Topics – We're All In This Together; From Theory to Practice

I. Introduction

The Digital Preservation Declaration of Shared Values, first put forth in 2017, sets aspirational but achievable standards for the efforts of the Digital Preservation Services Collaborative (DPSC) [1]. The DPSC is a volunteer alliance of representatives from community-supported digital preservation service providers, including APTrust, Chronopolis, CLOCKSS, LYRASIS, MetaArchive, and Texas Digital Library. These collaborating organizations are united in their commitment to preserve the cultural, intellectual, scientific, and academic record for current and future generations, using community-supported approaches. For the past six years, the values in the Declaration have established a foundation for these organizations to work together with a sense of trust and mission alignment.

Community-supported digital preservation initiatives foster community empowerment through governance, transparency, and accountability. These services also empower the organizations that are stewarding digital content to make informed decisions by providing them with much more than "black box" solutions. The organizations and the practitioners they serve, however, operate in a challenging technological and economic landscape where there are fewer and fewer resources for digital preservation. In the 2021 NDSA Staffing Survey, almost 70% of respondents stated that their

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

organization did not have the staffing needed to manage the digital content that they steward [2]. This has fostered a landscape where digital preservation functions are increasingly outsourced to commercial providers. Outsourcing these activities, without proper mechanisms for governance, transparency, and accountability, carries inherent risks for digital cultural heritage. Despite the benefits of values-centered approaches, however, community-based digital preservation service providers are operating in the face of claims that community-based approaches cannot innovate quickly enough to keep up with marketplace demands due to governance structures that are overly burdensome [3].

Within this neoliberal landscape, members of the DPSC desired closer and more intentional collaboration between their organizations, and they embarked on an IMLS Planning Project to propose a feasible service model for this collaboration [4]. The digital preservation community has long supported and recommended more collaborative approaches. The NDSA's 2015 National Agenda for Digital Preservation noted the need for a more coordinated ecosystem of distributed services [5]. This project is an effort to translate the DPSC's set of shared values into action, exploring exactly how much and what kind of collaboration among like-minded service providers is possible. In addition to finding strategic alignment and potential efficiencies between their services, the partners aim to demonstrate that the community governance and accountability offered by their services are not hindrances to innovation, but rather preconditions and catalysts for digital preservation good practice. The continued availability of services that prioritize transparency and accountability to the cultural heritage organizations they serve is necessary for these organizations to grant broad and sustained access to their digital material.

In this panel, three of the DPSC partners will discuss the importance of the shared values for digital preservation good practice, how they have enacted these values within their organizations, and other project findings.

II. Panel Format

This panel will explore the topic of operationalizing a set of community-supported and -supportive values into digital preservation practice. The format of the panel will be interactive, with panelists both reflecting on this topic and posing questions to the audience about their own digital preservation values and needs. Questions posed during the panel will include:

- How are these shared values being challenged?
- How can library executive-level staff contribute to value- and good practice-centric digital preservation programs?
- How can digital preservation service providers best support digital preservation staff in libraries, archives, and special collections?
- What are the risks of dependency on commercial providers of digital preservation services?
- How can non-commercial service providers incentivize community-based digital preservation partnerships?

III. Panelists

Hannah Wang is Program Officer for Digital Infrastructure at Educopia Institute, where she facilitates the work of the MetaArchive Cooperative and serves as Project Director for the DPSC Planning Project.

Jess Farrell is a Community Facilitator at Educopia Institute. She will moderate the session.

Courtney Mumma is the Deputy Director of the Texas Digital Library consortium, where one of her roles is managing Digital Preservation Services using Chronopolis and DuraCloud@TDL. She has worked in web archiving at the Internet Archive and is one of the creators of the Archivematica open source digital preservation workflow system.

Sibyl Schaefer is the Chronopolis Program Manager and Digital Preservation Librarian at the University of California, San Diego. She coordinates digital preservation activities across the UCSD Library and manages the Chronopolis distributed digital preservation system.

1. REFERENCES

[1] Digital Preservation Services Collaborative. 2018. "Digital Preservation Declaration of Shared Values." 2018. https://dpscollaborative.org/shared-values_en.html.

[2] National Digital Stewardship Alliance (NDSA). 2022. "2021 Staffing Survey," July. https://doi.org/10.17605/OSF.IO/EMWY4.

[3] Rieger, Oya, Roger Schonfeld, and Liam Sweeney. 2021. "The Effectiveness and Durability of Digital Curation Systems."

[4] "DPSC Planning Project: Sustainable Community-Owned Partnerships in Digital Preservation | Educopia Institute." n.d. Accessed March 9, 2023. https://educopia.org/dpsc-planning-project/.

[5] NDSA Agenda Working Group. 2014. "2015 National Agenda," September. https://osf.io/23vph/.

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

COMMUNITY IS WE

Modeling collective action as a framework for digital preservation

Alexandra Chassanoff
University of North Carolina Chapel Hill
United States
achass@unc.edu
0000-0002-1260-3031

Stacey Erdman
University of Arizona / Digital POWRR
United States
staceyerdman@arizona.edu
0000-0003-2569-5761

Jess Farrell
Software Preservation Network / Educopia Institute
United States
jess.farrell@educopia.org
0000-0001-9794-6908

Andrea Goethals
National Library of New Zealand
New Zealand
Andrea.Goethals@dia.govt.nz
0000-0002-5254-9818

Sharon McMeekin
Digital Preservation Coalition Scotland
Sharon.McMeekin@dpconline.org
0000-0002-1842-611X

Mikala Narlock
Data Curation Network/ University of Minnesota
United States
mnarlock@umn.edu
0000-0002-2730-7542

Hannah Wang
US National Archives and Records Administration
United States
hannah.wang@nara.gov
0000-0002-6676-1254

Abstract – This panel asks the question "how can collective action build global capacity for digital preservation?" Drawing on their own experiences participating in community-driven initiatives, panelists will describe and showcase how collective action efforts have created shared opportunities for advancing digital preservation goals. Following this discussion, panelists will reflect on the individual challenges and opportunities they faced in participating in such work. The panel will conclude with suggested next steps that can move the field globally towards a shared articulation of digital preservation work in practice.

Keywords – Best practices; education; collaboration; shared research

Conference Topics – We're all in this together; From theory to practice.

1. INTRODUCTION

Born-digital stewardship in contemporary GLAM settings presents information professionals (scholars, practitioners, educators, and students) with a multitude of ongoing, persistent and intersectional challenges. Born-digital materials, defined as "items created and managed in digital form" [1], carry inherent risk (of obsolescence, degradation, bit rot) and thus a more urgent timetable for preservation actions [2]. The frequency and speed of changes in the born-digital collecting sphere underscores the need to build community support mechanisms that provide collaborative environments for shared learning, troubleshooting, and skill building.

At the same time, community-driven efforts to address commonly experienced digital preservation problems through collective action have proven to be particularly effective in advancing practice. Collective action is broadly defined as measures taken by a group working toward a common objective. In the last decade, the formation and growth of many international, distributed

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

community networks have coalesced around the challenges and opportunities of current digital preservation work. They have created sustainable pathways for transitioning work from theory to practice by collaboratively developing resources and best practices, learning from the lived experiences of one another, and supporting rising professionals in learning the tools and skills to be successful. Showcasing the efforts of these communities has not been addressed comprehensively and efforts to map the terrain globally are underway [3].

2. PROPOSED PANEL

In response to the iPRES 2023 call for proposals related to the conference topic "We're All In This Together", this panel highlights the experiences of six community facilitators working in community-driven international digital preservation networks. Each panelist will be invited to contribute and reflect on topics such as the following:

- How does your community fit into the digital preservation landscape?
- How does your community define and approach "collective action"?
- How has your community collectively addressed a shared preservation challenge?
- How has your community balanced theory and practice?
- What challenges has your community faced in working collaboratively?
- Are there any upcoming projects or activities your community is hopeful to examine?
- How have you addressed sustainability in both your project and community formation?

The panel will be moderated to encourage both active discussion and audience participation.

3. INVITED PANELISTS

Invited panelists along with their affiliation and represented network (bolded) are described below:

Alexandra Chassanoff, Assistant Professor at the University of North Carolina at Chapel Hill, will moderate the panel.

Stacey Erdman is the Digital Preservation Librarian at the University of Arizona and currently serves as Project Director for the **Digital POWRR Peer Assessment Program**, and as an instructor for the Digital POWRR Institute training events.

Jess Farrell is a Community Facilitator for the **Software Preservation Network** at the Educopia Institute.

Andrea Goethals, Digital Preservation Manager at National Library of New Zealand, represents **Australasia Preserves**.

Sharon McMeekin is Head of Workforce Development with the **Digital Preservation Coalition**.

Mikala Narlock is the Director of the **Data Curation Network** based at the University of Minnesota.

Hannah Wang is the former Community Facilitator for the **MetaArchive Cooperative**.

4. REFERENCES

[1] Erway, Ricky. (2010). "Defining Born Digital." Report produced by OCLC Research. http://www.oclc.org/research/activities/hiddencollections/borndigital.pdf

[2] AIMS Work Group. (2012). AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship. http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final.pdf

[3] Lindlar, Michelle; Pohlkam, Svenia; Zarnitz, Monica; Bahr, Thomas; Strathmann, Stefan. (2022). "Mapping the Landscape of Digital Preservation Networks." Proceedings of the Eighteenth International Conference on Digital Preservation (iPRES), Glasgow, UK, September 12-16, 2023.

LESSONS FROM THE FUTURE

Looking Back on Policy Development

Elizabeth England
U.S. National Archives and Records Administration
USA
elizabeth.england@nara.gov
0000-0002-6432-8123

Martin Gengenbach
National Library of New Zealand
New Zealand
martin.gengenbach@dia.govt.nz
0000-0003-2180-6727

Sharon McMeekin
Digital Preservation Coalition
Scotland
sharon.mcmeekin@dpconline.org
0000-0002-1842-611X

Jenny Mitcham
Digital Preservation Coalition
UK
jenny.mitcham@dpconline.org
0000-0003-2884-542X

Kieran O'Leary
National Library of Ireland
Ireland
koleary@nli.ie
0009-0009-7485-0634

Abstract – Policy is an important component of a successful digital preservation program. For example, CoreTrustSeal [1] suggests that a policy statement would be appropriate evidence to demonstrate that a repository has an explicit mission to provide access to and preserve digital objects, and the DPC's Rapid Assessment Model [2] suggests that a digital preservation policy should be in place in order to reach the 'Basic' level of the 'Policy and Strategy' section. While resources exist [3] to assist organizations in developing their first digital preservation policy, these formative strategic documents are intended to hold relevance beyond their initial publication. This panel session highlighted challenges and opportunities in the development and ongoing maintenance of digital preservation policies across three organizations: U.S. National Archives and Records Administration, National Library of Ireland, and National Library of New Zealand. Panelists reflected on learnings from different stages of the policy lifecycle, including initial development, initiating revisions, and re-engaging with dormant policy documents. These efforts are contextualized within broader policy education resources, including the DPC's revised Digital Preservation Policy Toolkit [4].

Keywords – policy, outreach, documentation, advocacy

Conference Topics – From Theory to Practice; We're All in this Together

I. INTRODUCTION

Digital preservation policies represent many things to many organizations. For some, publication of a policy represents a foundational event in a digital preservation program; for others it is an aspirational document that guides developing operations; for others it indicates a level of operational maturity and stability. For many organizations, it serves more than one of these roles.

Because digital preservation policies are so specific to an organization and its setting, it can be challenging to transfer generalized policy guidance to a unique organizational context. This panel was put together to explore the challenges in developing and maintaining digital preservation policies across different stages of the policy lifecycle, drawing lessons learned and recommendations from practitioners across the world: from those contemplating their first policy to those who may have inherited a policy that no longer meets the needs of their organization.

II. THE PANELISTS

The panelists for this submission were selected for their diverse policy experiences; a short description of each panelist and their work in policy development is provided below. This session was organized in collaboration with Jenny Mitcham and facilitated by Sharon McMeekin, both of the Digital

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

Preservation Coalition (DPC). The DPC have revised and republished their Digital Preservation Policy Toolkit this year and are developing training materials on the topic of digital preservation policy development.

Elizabeth England is Senior Digital Preservation Specialist at the U.S. National Archives and Records Administration (NARA), where she participates in strategic and operational initiatives and services for the preservation of born-digital and digitized records of the U.S. federal government. The NARA digital preservation strategy was first published in 2017 as a largely aspirational document, and Elizabeth led revisions to the document in 2022.

Martin Gengenbach is Digital Preservation Policy and Outreach Specialist at the National Library of New Zealand (NLNZ). His role is focused on developing and communicating policy to support digital preservation throughout the Library. He began this role in 2022, and has been driving revisions to the Library's digital preservation strategy and digital preservation policy, which were originally published in 2012.

Kieran O'Leary is Digital Preservation Manager at National Library of Ireland (NLI). He is responsible for coordinating the implementation of digital preservation throughout the Library. NLI has drafted previous digital preservation policies in 2017 and 2020, and will publish their first digital preservation policy in 2023.

III. Panel Discussion Topics

During this session, panelists reflected on what they have learned in their work developing, maintaining, and updating digital preservation policies. The topics and questions chosen for discussion constitute "lessons from the future," providing guidance for policy development rooted in past experience and common challenges.

Building internal support for policy development - Administrators and funders may not understand the purpose and value of a digital preservation policy, particularly if there are already operational procedures in place for working with digital materials. Panelists discussed their experiences in cultivating support for policy publication and revision, and where new internal champions were needed to move forward with policy goals.

At NARA it was noted that a gap analysis based on ISO 16363 was carried out in 2017 and identified the need to have a digital preservation policy in place. Knowing that at that stage, the document would be largely aspirational, it was eventually agreed with colleagues that framing it as a strategy would be more appropriate than having a policy document that NARA didn't meet in practice. It was noted however that having a document (even if not policy in name) was better than not having one at all.

At NLI there has been a lot of support for the development of a digital preservation policy and to a certain extent this has been supported by the presence of digital preservation in the institutional risk register. Once the risks around digital preservation are flagged up to senior management it becomes more pressing to find ways to mitigate them. This was a key step in gaining necessary buy in and support from colleagues. Another important step was to form a steering group of key stakeholders who could review and have oversight of the policy. At NLI an existing group was transformed and repurposed and this has been key to moving the policy forward.

Martin Gengenbach noted that at NLNZ support can take many different forms. The very fact of his job role being supported suggests that policy is a priority within the organization. Challenges around gaining support have been in getting individual units within the organization to engage with policy work around other operational priorities. He has discovered that the 'outreach' aspect of his job title is equally important to the 'policy' element, with the two parts going very much hand-in-hand.

How policy can be aligned with organizational strategy and vision - Connecting a digital preservation policy to organizational priorities is one way to gain administrative support by demonstrating how digital preservation goals further other organizational initiatives. Panelists were encouraged to discuss some of the broader organizational strategy elements that played a role in their digital preservation policy development.

At NLNZ digital preservation policy work has been tied in to ongoing initiatives that directly impact business units across the organization. It has been

incredibly valuable to be able to demonstrate how digital preservation policy work is applicable to wider organizational goals and initiatives.

At NARA, Elizabeth England expressed an intention to firmly tie policy revision into the timeline of the organizational wide strategic plan. The current plan runs until 2026 which marks a key change in digital collecting for the organization. Digital preservation policy needs to align with, and support, this wider plan. Review cycles for these documents will also be aligned in future.

Kieran O'Leary noted a similar situation with policy aligning with wider organizational strategy. It is early days for the new digital preservation policy at NLI and anticipated that an annual review may be necessary initially, but that it may be possible to align policy review with the five-year strategic planning cycle in the future.

Communicating policy, internally and externally - How policy is communicated both internally and externally will have an impact on implementation. Panelists explored different communication strategies and their efficacy in their unique organizational contexts, highlighting the need for ongoing communication throughout the development process to ensure all stakeholders remain informed and engaged.

Communication is a key part of Martin Gengenbach's role at NLNZ. As a relatively new employee, his first year has largely been about communication – talking to key stakeholders and finding out what their challenges are, as well as understanding how policy has been created and maintained in the library in the past. He noted that it is OK to over-communicate and that providing multiple opportunities for comment and feedback is not a bad thing. He recognizes that whilst he may think about policy all the time, other stakeholders within the organization are being pulled in many different directions. Frequent communication in a number of different ways helps to keep policy in their minds.

Elizabeth England described how NARA keeps stakeholders informed and engaged through their digital preservation guidance group. This group includes representatives from across the organization, including the custodial units responsible for records received from three different areas of government that supply records to NARA. Having input from all of these different areas (all with different regulations) helps to keep the digital preservation strategy broadly relevant to all stakeholders.

Framing policy as present state or aspirational - Depending on the existing state of digital preservation operations, the digital preservation policy may be framed as an aspirational statement of intent ("we will"), or an articulation of current practice ("we do"). Panelists shared their perspectives on the factors that impact how an organization may choose to frame their policy.

The most recent digital preservation policy developed at NLI was intended to reflect present state, but external feedback given on an early draft highlighted that use of the future tense in policy statements led to it being misunderstood as aspirational. This issue has been resolved in its latest version. Whilst most of the policy reflects current state, there are a few areas within the policy that mention areas of work that will be developed in the future. NLI plans to implement annual check-ins using DPC's Rapid Assessment Model and ensure that continuous improvement is at the heart of their digital preservation work.

Martin Gengenbach's initial impressions of the existing NLNZ policy manual when first encountering it was that it reflected the present state. In this case, the policy statements were very specific and granular. In actual fact, it had been developed as the organization tried to understand how they would use their digital preservation repository rather than based on processes that were actually operational. The rewrite of this policy will more closely reflect the fully functional digital preservation program and will aim to be present state. It was also noted that the new policy will be higher level, leaving out much of the procedural detail which is more suited to other forms of documentation.

Elizabeth England noted that the original 2017 strategy was deliberately aspirational. It had been developed after a gap analysis was carried out, and the policy was very much intended as a way of committing to bridge those gaps that had been identified. In the more recent revision of this policy, many of those aspirational statements now reflect the current state. The updates made to the policy

included reframing the language to use "we do…" instead of "we will…".

Turning aspiration into operation - Ensuring the successful implementation of a policy demonstrates accountability and builds trust in an organization's digital preservation program. Panelists were asked to reflect on how new policy can support existing procedure; where policy and implementation combined can identify and resolve gaps in current practice; and how thoughtful implementation can support later policy goals.

The NLI policy has an implementation and next steps section which is aimed to help move any aspirational goals forward. The close alignment of their policy with DPC's Rapid Assessment Model has also helped with highlighting concrete steps that could be taken to improve, and the planned yearly cycle of RAM assessment will continue to move this forward over time.

At the NLNZ the current priority is to make sure the policy is in alignment with current operations.

Elizabeth England described a "push versus pull" between policy influencing practice versus practice influencing policy. She noted that her revision process includes creating documentation about elements that have moved from aspiration to operation.

Lessons from the future – This panel discussion was all about lessons from the future and the panelists have clearly all learned much from their work in this area. They were asked to summarize the key messages they would pass back to their past selves at the beginning of their digital preservation policy journeys.

Martin Gengenbach noted the importance of setting goals from the outset. He stressed the benefits of ensuring that you have a clear understanding of why a change to policy might be necessary in your specific context. Reviewing these goals regularly is also key.

Elizabeth England chose to flag up the value of documentation. Documenting the process of creating or revising your policy or strategy will be a huge help to your future self. Recording why you made particular decisions, why you worded something in a particular way, and of course, whether your policy is aspirational or present state will be incredibly helpful to anyone who comes to revise it.

Kieran O'Leary recognized the value of engaging all relevant stakeholders as early as possible in the policy creation or review process. Having a steering group with all the right people around the table was hugely beneficial to the work on preservation policy at NLI.

IV. Conclusion

While generic good practice guidance (such as that found in the DPC's Digital Preservation Policy Toolkit) can be helpful for those who are getting started with writing or reviewing policy, it is also helpful to hear the experiences of different organizations who have tackled this challenge. There is no one-size-fits-all approach to preservation policy and each organization must find a unique approach to meet their own needs. This panel session provided an opportunity to learn about how this task was approached in practice and to discuss key themes across different contexts, highlighting both contrasting approaches and parallels.

1. REFERENCES

[1] CoreTrustSeal Trustworthy Digital Repositories Requirements 2023-2025 V01.00, 2022 https://doi.org/10.5281/zenodo.7051011

[2] Rapid Assessment Model, Digital Preservation Coalition, Version 2.0, 2021 http://doi.org/10.7207/dpcram21-02

[3] Digital Preservation Policy Toolkit Further Resources, Digital Preservation Coalition. Version 2.0, 2023 https://www.dpconline.org/digipres/implement-digipres/policy-toolkit/policy-resources

[4] Digital Preservation Policy Toolkit, Digital Preservation Coalition. Version 2.0, 2023 https://www.dpconline.org/digipres/implement-digipres/policy-toolkit

TIPPING POINT

Have we gone past the point where we can handle the Digital Preservation Deluge?

Paul Stokes
Jisc
UK
paul.stokes@jisc.ac.uk
0000-0002-7333-4998

Karen Colbron
Jisc
UK
karen.colbron@jisc.ac.uk
0000-0002-2438-4008

Abstract – The world today is faced with an insurmountable problem. There is too much digital "stuff" in existence for us to even handle in any sort of meaningful way, let alone curate and preserve. We have reached (or perhaps even gone beyond) the data processing tipping point. There is an enormous amount of data already in existence and unimaginably more being generated every day. This panel proposal (and accompanying poster) is intended to explore this doomsday data scenario with a group of experts in the field of Digital Preservation and related disciplines with a view to deciding if it is true and what can be done about it.

Keywords – Data doomsday, Tipping point, Data deluge

Conference Topics – SUSTAINABILITY: REAL AND IMAGINED;. WE'RE ALL IN THIS TOGETHER; IMMERSIVE INFORMATION

I. INTRODUCTION

Some facts about the data-verse we currently inhabit.

There is an enormous amount of data existing in the world today. According to the International Data Corporation (IDC), the amount of digital data created, captured, and replicated in 2020 was approximately 64.2 zettabytes (1 zettabyte = 1 trillion gigabytes)[1]. This figure is expected to grow to 181 zettabytes by 2025, which represents a compound annual growth rate of 23%[1]. It is interesting to note that this figure is constantly being revised upwards. Publications from as recently as the late 2010's had this figure estimated as just over half that figure.

A report/infographic from the World Economic Forum based on data from by Seagate and IDC found that the amount of data generated each day is expected to reach 463 exabytes (1 exabyte = 1 billion gigabytes) by 2025, up from 23 exabytes in 2018[2]. This represents a CAGR of 29.4% over the seven-year period. The rate of world production of digital data is increasing at an extremely fast pace, with the amount of data generated each year growing by tens (possibly even hundreds in the future) of percent.

Why is this something to be worried about? Well generating the data is just beginning of the potential problem. The data needs to be transported, copied, and stored (and in some cases, curated and preserved), all of which require resources (including power) that are finite… and not keeping pace.

More facts.

As of 2022, it was estimated that the world had approximately 6.7 zettabytes (ZB) of data storage capacity and that this would rise to around 16 zettabytes by 2025 according to statistia[3]. This estimate includes all types of data storage, including hard disk drives, solid-state drives, optical storage, and tape storage. This number is constantly increasing, but not at the same pace as data production. Yes, there are newer and denser storage media on the horizon (or even in production) such as DNA and storage on atoms, but even that is finite. An estimate published in the Straits Times gave a figure of about 180 years before all the atoms on earth were used for storage at the current rate of data production[4]. This is quite clearly a nonsense scenario, but it gives a flavour of the problem.

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

And what about manipulating the data. Ingesting it into a digital preservation system for instance. Doing it manually at high speed and large volumes is out of the question*. Semi-automated ingest (possibly enhanced by AI) should be faster. However, even this style of ingest is clearly several orders of magnitude slower than the rate at which data is being produced. Anecdotally, a growing number of data stewards are reporting that they are receiving data deposits faster than they can process them. As a result, the unprocessed data is being put into (at best) bit preserved long-term storage for processing "at a later date…" a later date that keeps moving further away. Automated ingest helps, but is still not going to keep pace.

So have we reached the tipping point? Are we past the point at which we can meaningfully process the deluge of data being generated? Is there anything that can be done to mitigate against data doomsday? That's what we'd like to explore in this panel.

II. The panel

We see this as an interactive panel session utilizing the expertise of up to 7 authorities from the field of Digital Preservation and related disciplines. Each will be asked to briefly put forward a point of view / opinion relating to the veracity (or otherwise) of the data doomsday scenario. The statements will be followed by a series of questions designed to explore how the situation could be either avoided or mitigated.

The audience will be invited to participate through a series of interactive polls and questions/observations from the floor (both those in attendance and those attending virtually). In particular, the audience will be polled at the beginning and the end of the session to see if they have been persuaded to shift their pre-extant opinions regarding the data doomsday scenario by the arguments presented in the session.

The following panelists have expressed a willingness to take part:

- William Kilbride—Executive Director of the DPC
- Matthew Addis—Chief Technology Officer at Arkivum
- Stephen Abrams—Head of Digital Preservation at Harvard University
- Kate Murray—Digital Projects Coordinator at Library of Congress
- Nancy McGovern—Director of Digital Preservation at MIT Libraries
- Tim Gollins—Director of Vanderbilt University Special Collections and University Archives
- Helen Hockx-Yu—Enterprise Architect at University of Notre Dame

1. REFERENCES

[1] International Data Corporation (IDC). (2020). The digitization of the world from edge to core. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[2] https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/!

[3] Total installed based of data storage capacity in the global datasphere from 2020 to 2025, https://www.statista.com/statistics/1185900/worldwide-datasphere-storage-capacity-installed-base/

[4] https://www.straitstimes.com/singapore/world-faces-data-storage-crunch-ahead

* A moments consideration of the wide number and variety of processes and systems involved in the ingest process (multiple carrier types, multiple file types, multiple processes on multiple different infrastructures) leads us to the conclusion that a definitive, quantitative measure of manual ingest rate is impractical.

RETROSPECTIVE, SUBJUNCTIVE, PROSPECTIVE: PROVENANCE CHALLENGES ACROSS TIME

Rhiannon Bettivia
Simmons University
USA
bettivia@simmons.edu
0000-0003-4593-562X

Yi-Yun Cheng
Rutgers University
USA
yiyun.cheng@rutgers.edu
0000-0001-6123-7595

Michael R. Gryk
University of Illinois
USA
gryk2@illinois.edu
0000-0002-3483-8384

Abstract – This panel will explore provenance: as theory and practice; as a tool for sustainability; and as a space of shared struggle and challenge for digital preservations and those in fields ranging from archives to cluster computing. In digital preservation, provenance tells us where an object has come from, the myriad preservation actions we could take to care for it, and where we predict the object will need to go in future. This panel is intended for anyone who is interested in the world of provenance: defining it, understanding it, modeling it, addressing the vague dissatisfaction practitioners often have when researching and documenting it. Provenance is more about the journey than the destination: this panel aims to surface a variety of experiences with provenance and to facilitate discussion and a community of practice around the relationship between digital preservation and provenance.

Keywords – provenance, authenticity, evidence, archival values, information theory

Conference Topics – SUSTAINABILITY: REAL AND IMAGINED; WE'RE ALL IN THIS TOGETHER; FROM THEORY TO PRACTICE

I. INTRODUCTION

The OED defines "provenance" as origin, source, and ownership, tied tightly to the ability to determine authenticity. Provenance can be used to describe what did happen (retrospective provenance), what could happen (subjunctive provenance), and what will happen (prospective provenance). Provenance has many faces in different fields: the tree of life in phylogeny; ancestry of families in genealogy; layers of sediments in stratigraphy. Provenance transcends disciplines. In digital preservation, the custodial chain, audit trails, iteration reports, and change logs are building blocks for establishing authenticity in the face of managed change over time. Measurable properties and questions about the identity of digital objects engender challenges in modeling and recording different stages of computational projects. The lack of provenance information for born-digital objects in each stage of a research pipeline can reduce the transparency, trustworthiness, and reproducibility. Xu et al. [6] state that reliance on process has actually changed and expanded traditional uses of the term provenance:

"The notion of provenance has been adopted and extended in the field of Computer Science and applied to concepts such as data, computation, user interaction, and reasoning. In this context, provenance is no longer limited to origin or history, but also includes the process and other contextual information."

Technologies like blockchain are bound up with procedural provenance in their very form and function [5,3]. Provenance stories will play an increasing role as AI artifacts may impede the ability of archival materials to accurately represent the historical records [4].

We define provenance broadly as how something has come to be, and we incorporate the following key concepts into our exploration of provenance [1,2]:

- Provenance is fluid and transcends time;
- Creating provenance descriptions is both a conceptual modeling, a metadata recording exercise, and a persuasive exercise;
- Working with provenance is both a ubiquitous and field-agnostic act.

As they became more commonly encountered in archives and other information institutions, digital

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

records destabilized commitments and assumptions of traditional preservation. The lack of provenance information for born-digital objects in each stage of a research pipeline can reduce the transparency, trustworthiness, and reproducibility. Archival concerns over the mutability of digital records eventually gives way to a realization that their particular affordances may support additional techniques for ensuring provenance than physical records, such as embedded metadata, blockchain technologies, and digital forensics approaches.

Provenance becomes an intellectual and moral concern as collections of digital objects are managed through their life cycles, migrated, emulated, and remediated in new formats and interfaces, such as virtual and augmented reality. How does our conceptualization of provenance adapt to these new conditions and ensure that we can continue to trust the authenticity and integrity of copies over time?

II. Panel Structure

We have assembled a panel to have an exploratory discussion about the concept of provenance. This 90-minute panel brings together participants from across the ASIS&T community to represent concerns from information organization, research data management, metadata, cultural heritage, archives, digital curation, data curation, and digital preservation.

A. Part I: Establishing a Baseline

B. Each panelist will share their thoughts on provenance and how it intersects with their work.

C. PART II: Interactive Q&A

The panel will address topics posed by a moderator. These topics include:

-PREMIS as a provenance standard

-Workflows as prospective provenance

-Subjunctive provenance as a mediator with future audiences

III. Panelists and Coordinators

Karin Bredenberg (invited panelist) is the Metadata Strategist at the Kommunalförbundet Sydarkivera, a local federation of 37 municipalities in Sweden.

Dr. Alexandra Chassanoff (invited panelist) is an Assistant Professor at the School of Library and Information Sciences at North Carolina Central University.

Dr. Zack Lischer-Katz (invited panelist) is Assistant Professor in Digital Curation and Preservation at University of Arizona's School of Information.

Dr. Mike Twidale (invited panelist) is a Professor and the PhD Program Director at the iSchool at the University of Illinois, Urbana-Champaign.

Dr. Rhiannon Bettivia (coordinator) is faculty at Simmons University in Library and Information Science.

Dr. Yi-Yun (Jessica) Cheng (coordinator) is faculty at SC&I at Rutgers University.

Dr. Michael R. Gryk (coordinator) is Associate Professor of Molecular Biology and Biophysics at UCONN Health.

1. REFERENCES

[1] Bettivia, R., Cheng, Y., & Gryk, M. (2022a). Documenting the Future: Navigating Provenance Standards. Springer Nature.

[2] Bettivia, R., Cheng, Y., & Gryk, M. (2022b). Storied Past, Bright Future: A Prov Jam Session. ASIS&T 2022 Annual Meeting, Pittsburgh, PA.

[3] Dell, N., Perrier, T., Kumar, N., Lee, M., Powers, R., & Borriello, G. (2015, February). Digital Workflows in Global Development Organizations. In Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (pp. 1659-1669).

[4] Prelinger, R. (2021, April 28). NFTs and AI Are Unsettling the Very Concept of History. Wired.

[5] Woodall, A., & Ringel, S. (2020). Blockchain archival discourse: Trust and the imaginaries of digital preservation. new media & society, 22(12), 2200-2217.

[6] Xu, K., Ottley, A., Walchshofer, C., Streit, M., Chang, R., & Wenskovitch, J. (2020, June). Survey on the analysis of user interactions and visualization provenance. In Computer Graphics Forum (Vol. 39, No. 3, pp. 757-783).

VOLUMETRIC VIDEO FOR PRESERVATION

Exploring the Possibilities and Challenges for Immersive BIPOC Storytelling

Zack Lischer-Katz
University of Arizona
USA
zlkatz@arizona.edu
0000-0002-4688-1275

Rashida K. Braggs
Williams College
USA
rkb2@williams.edu

Bryan Carter
University of Arizona
USA
bryancarter@arizona.edu

Sven Bliedung von der Heide
Volucap GmbH
Germany
sbliedung@volucap.de

Abstract – This panel explores the possibilities and challenges of volumetric video capture for digital humanities research and pedagogy, particularly in terms of documenting and representing the stories of BIPOC Americans who have lived through historical eras of global conflict. The panel will focus on the panelists' experiences working with volumetric video and their work on a multi-institutional National Endowment for the Humanities-funded project. Panelists will offer perspectives on the benefits of volumetric video and its preservation challenges.

Keywords – volumetric video, immersive media, digital storytelling, inclusion, preservation

Conference Topics – Digital Accessibility, Inclusion, and Diversity; Immersive Information

I. INTRODUCTION

Volumetric video techniques offer new possibilities for immersive storytelling, producing an experience of realistic presence of people and objects in augmented reality (AR) and virtual reality (VR). This emerging technology has the potential to transform research and teaching for a range of areas, including cultural heritage preservation, oral histories, and spatial understanding of historical spaces and people [1, 2].

Volumetric video is one of the most recent media formats to be considered for use by digital humanities (DH) scholars, historians, and digital media producers. It has emerged at a time when VR, AR, and XR (extended reality), are becoming increasingly affordable for scholarly use and cultural heritage preservation [3, 4, 5]. Volumetric capture produces 3D content that has a time-based, cinematic dimension. Each frame of volumetric video is a 3D model of the subject, which enables full rotation and viewing of the subject in space [6], enhanced presence of the subject and engagement for a variety of potential users and applications. Using an array of multiple depth-sensing cameras arranged around the subject, volumetric video captures visual and depth data. The resulting assets can be integrated into XR environments [7]. Investigating volumetric capture from a DH perspective entails both exploring its possibilities for humanities research, digital storytelling, and cultural heritage preservation, as well as interpreting how volumetric video, integrated as it is into other media, such as feature-length Hollywood films or the AR apps on our smartphones, shapes our lived experiences in the 21st century.

This panel will focus on panelists' experiences with volumetric video and their partnership on a multi-institutional project funded through a grant from the National Endowment for the Humanities (NEH) - Digital Humanities Advancement Program

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

(with University of Arizona, Williams College, and the company Volucap, GmbH). The project team is exploring how volumetric video can be used to uniquely preserve narratives and cultural memories of BIPOC (Black, Indigenous, & People of Color) World War II era American veterans, as well as developing best practices for capturing and preserving more inclusive digital histories.

II. Panel Structure

Each of the four panelists will speak briefly about their role in the project and their experiences working with volumetric video as an immersive medium. The audience will be brought into the conversation through a moderated discussion with panelists, guided by audience-supplied questions.

A. Volumetric Video for Immersive Digital Humanities Storytelling (Dr. Bryan Carter)

Volumetric video technology is still very expensive; however, recent prosumer level hardware and software now make it possible for humanities researchers with a medium-sized budget to use. This talk explores the hardware, software and knowledge base necessary to make use of volumetric video capture for digital storytelling.

B. Connecting Black World War II Memories to Black Futures through Volumetric Video Capture (Dr. Rashida K. Braggs)

This talk will consider questions, insights and challenges that have arisen in interviewing African American WWII veterans and family members for this digital storytelling project, asking: Which narratives will resonate most with young American students in danger of their multicultural histories being erased from their curricula? What are best practices for ensuring authentic representation of their stories? How can immersive technologies be used to explore these questions?

C. Volumetric Video Capture from the Film Industry Perspective (Sven Bliedung von der Heide)

Sven Bliedung von der Heide will discuss Volucap's work on *The Matrix: Resurrections* and the narrative possibilities for volumetric technology. Volucap is known for its volumetric studio in Potsdam, Germany, where it has developed novel approaches to capture cinema-quality 3D images of actors moving on real sets. Applications also lie in new forms of interactive storytelling for representing history in immersive and engaging ways.

D. Challenges of Curating and Preserving Volumetric Video (Dr. Zack Lischer-Katz)

Digital curation and preservation guidelines are still being developed for volumetric video. This talk extends recent research on the preservation challenges of VR and 3D data [8, 9, 10] to explore its digital preservation and curation challenges, including file formats, appraisal and selection criteria, legal and ethical issues, and repositories.

III. Conclusion

Volumetric video is being "democratized" through decreasing costs and increasing use in humanities research [1]. By starting a discussion in the preservation community, this panel hopes to encourage further research on best practices for the curation and preservation of volumetric video.

IV. REFERENCES

[1] Carter, B., et al. (2021). The Robinhood of 3D: Democratizing volumetric capture for more accessible and innovative learning environments. In *ICERI2021 Proceedings* (pp. 4845-4850). IATED.

[2] Gamber, C. (2022, July). Encountering Pinchas Gutter in virtual reality and as a "hologram": Immersive technologies and one Survivor's story of the Holocaust. In *Intelligent Computing: Proceedings of the 2022 Computing Conference, Volume 1* (pp. 358-374). Springer.

[3] Lischer-Katz, Z., et al. (2019). Introduction–3D/VR creation and curation: An emerging field of inquiry. *3D/VR in the Academic Library: Emerging Practices and Trends, CLIR Report 176*, https://www.clir.org/pubs/reports/pub176/

[4] Wang, J., & Lu, C. (2022, August). Research on the development and practice of digital technology in architectural heritage. *2022 International Conference on Culture-Oriented Science and Technology)* (pp. 146-150). IEEE.

[5] Verschure, P. F., & Wierenga, S. (2022). Future memory: a digital humanities approach for the preservation and presentation of the history of the Holocaust and Nazi crimes. *Holocaust Studies*, *28*(3), 331-357.

[6] O'Dwyer, N., et al. (2021). Volumetric video in augmented reality applications for museological narratives: A user study for the Long Room in the Library of Trinity College Dublin. *Journal on Computing and Cultural Heritage (JOCCH)*, *14*(2), 1-20.

[7] Schreer, O., et al. (2022). Preserving memories of contemporary witnesses using volumetric video. *i-com*, *21*(1), 71-82.

[8] Campbell, S. (2017). *A Rift in our Practices, Toward Preserving Virtual Reality*, Masters Thesis, New York University.

[9] Lischer-Katz, Z. (2020). Archiving experience: An exploration of the challenges of preserving virtual reality. *Records Management Journal*, *30*(2), 253-274.

[10] Moore, J., Rountrey, A. and Kettler, H.S. (eds.). (2022). *3D data Creation to Curation: Community Standards for 3D Data Preservation.* Association of College and Research Libraries.

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

DIGITAL ACCESSIBILITY, INCLUSION AND DIVERSITY

Digitization of Indigenous Agricultural Knowledge in Shaping Food Security across the Kenyan Coastal Region

Maureen Kenga

Pwani Ufanisi Farmers' Cooperative
Kenya
Maureen.kenga@gmail.com
https://orcid.org/0000-0001-6683-4495

Abstract – According to Gilman (1917), food problem is related to three questions: First, "how to produce the most food with the least cost in time, labor and money," second, how to swiftly, efficiently and economically distribute it to consumers, and, third, how to prepare and serve healthy food, without spending too much money, time, and effort. Since then, considerable progress has been made in improving food supply and facilitating meal preparation. This paper looks at the importance of digitizing the indigenous farming methods that can be incorporated with the emerging trends in present agriculture playing a significant role in ensuring food security. This paper is submitted to iPRESS 2023 as a poster aimed at improving indigenous agricultural knowledge with a base for new capabilities in providing solutions matters food security. The paper addresses the conference topic: Digital accessibility, Inclusion, and Diversity.

I. INTRODUCTION

African Indigenous Foods Systems were clear and properly designed to ensure that households fed themselves. In the Kenya coastal region, many activities took place to ensure that families were food secured and this include: Shifting cultivation on food crops, cash crops and horticulture crops; Intercropping involving annual (mostly food crops) and perennial crops (coconut and cashew nuts); Rain water harvesting to water crops; Farming activities like digging, planting, weeding and harvesting done by both men women and children (i.e. family); and preserving harvest for use before the next season using traditional methods like wooden barns (Were, 1988). Indigenous knowledge of preserving cereals, vegetables and meat were applied to ensure that there was enough food even during drought season. All these ensured food security at house level. With the current climate change issues and recent draught experienced in Kenya for the past six years, some indigenous methods of agriculture can play a crucial role in reversing the problems. This can only happen if the indigenous methods are digitized and shared with an aim of transforming food security.

Usually, farmers, either on their own or cooperatively in conjunction with other members of their neighborhoods, develop their respective knowledge base through time as a product of their interaction with the environments in which they practice their livelihoods. Knowledge and skills are derived through a system of experimentation, spatial cognition and perceptions that lead to the selection of the most adaptive and useful practices. Successful adaptations and practices are preserved and passed on through generations mainly through oral tradition and on-farm practice.

A. Indigenous Agricultural Knowledge

It has become essential for scientists to comprehend traditional agriculture and the knowledge base that it offers due to the ongoing production issues facing crops and livestock, such as the regular crop failures brought on by drought, flooding, and insect infestations. It is clear that a complex farming system has improved traditional farmers' grasp of their

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

surroundings, cropping and livestock movement networks, and helped them manage severe conditions to meet their subsistence needs without relying on contemporary agricultural technologies. The shortcomings that modern agriculture is currently experiencing might be fixed by comprehending these knowledge systems.

Indigenous agricultural knowledge include:
1) Well-established calendars for crops and livestock movement, productivity linkages between soil and drainage, climatic changes, and the function of natural plants and wildlife as environmental vitality indicators.
2) Farmers' knowledge of the environment profoundly influenced their decision-making about the site and timing of their produce.
3) Inter-cropping, which involves farmers having a thorough understanding of the types of plants, animals, insects, and birds that can or cannot coexist, as well as the functions of insects and other related arthropods as crop pests, disease-causing agents, food sources, and medicinal agents in their production systems. Inter-cropping increases production and guarantees food security.
4) Knowledge of environmental factors, including fauna and flora, leads to proper farming techniques in the face of floods, droughts, pests and diseases, and low soil fertility, which improves their ability to cope.
5) Traditional farmers mix a large number of species with structural variety throughout time and location (both through vertical and horizontal organization of crops). Some crops act as supports or provide shade to others by being grown together.
6) Farmers take full advantage of the variety of micro-environments present in a field or region, which include those with varying soil, water, temperature, height, and slopes.
7) Crop-livestock alliances, in which livestock graze in vacant fields and leave manure behind, are another typical method used to preserve soil fertility and guarantee the availability of cropland throughout the growing season.
8) The traditional crops are high nutrition foods that are eaten at home or traded locally. Many of these crops along with numerous wild plants are also used medicinally. What might be considered a weed in some communities is often eaten as a salad by others. Every house hold lists several unique medical plants

B) Digitization of Indigenous Agricultural Knowledge

At the farm level and throughout the value chain, digitizing agricultural knowledge improves efficiency, productivity, and sustainability (Aubert et al., 2012; Wolfert et al., 2017). Agriculture information is being digitized in order to preserve it for future generations and to lessen the difficulties that farmers are having with contemporary agriculture. Digital platforms and applications have the power to fundamentally alter how information is processed, shared, accessed, preserved and used. Digital applications will enable hitherto impractical decision-making for farmers, potentially resulting in fundamental changes to farm management (Sonka, 2014; Wolfert et al., 2017).

The team has come up with a project of identifying this agricultural knowledge in the Kenyan coastal region and digitize it. The project is as a result of ideas that cropped up during the cooperative's education days and farm field visits where farmers shared their agricultural knowledge thus the need for digitization and preservation for easy sharing and retention. So far, the team has formed a farmers' cooperative (Pwani Ufanisi Farmers' Cooperative Society – PUFCO), that currently has slightly over six hundred farmers clustered depending on the crops that they farm and possess the crucial indigenous agricultural knowledge needed for digitization. A community of practice group has been formed via a WhatsApp group whereby the farmers share knowledge freely. Physical farm visits are usually organized for physical knowledge transfer out of which some have been digitized. Several indigenous seeds including maize, sesame, cassava cuttings, and cashew nuts have been collected and are being reproduced by various farmers with the guidance and professional help from Kenya Plant Health Inspectorate Service (KEPHIS). PUFCO has started the process of producing documentaries on the various types of indigenous agricultural knowledge. The farmers are also trained on how to take videos of the various farming activities they are undertaking in their farms and how they are applying indigenous agricultural knowledge. The photos, videos and WhatsApp chats are exported into the PUFCO

website (www.pwaniufanisi.co.ke) digital repository for preservation and where members can access, share and learn.

C. Conclusion

Digitization of indigenous agricultural knowledge and sharing the same to various stakeholders will play a crucial role in in ensuring sustainable food production and supply in Kenya and other parts of Africa. Great knowledge will also be shared on various ways of improving traditional crops as sources of income thus eradicating poverty. Documenting, digitizing and preserving indigenous agricultural knowledge will also add value in understanding the nutritional value of respective crops, combating climate change issues, promoting health, education and promoting sustainable consumption and production. Foreseen challenges on this project will be expertise, funding and lack of equipment.

1. REFERENCES

[1] Aubert, B. A., Schroeder, A., and Grimaudo, J. (2012). IT as enabler of sustainable farming: an empirical analysis of farmers' adoption decision of precision agriculture technology. Decis. Support Syst. 54, 510–520. doi: 10.1016/j.dss.2012.07.002

[2] Gilman, Charlotte. (1917). the Housekeeper and the Food Problem. Annals of the American Academy of Political and Social Science - ANN AMER ACAD POLIT SOC SCI. 74. 123-130. 10.1177/000271621707400118.

[3] Sonka, S. (2014). Big data and the ag sector: more than lots of numbers. Int. Food Agribus. Manag. Rev. 17, 1–20. doi: 10.22004/ag.econ.163351

[4] Were, Gideon. (1988). Kilifi District Socio-Cultural Profile. University of Nairobi. Nairobi

[5] Wolfert, S., Ge, L., Verdouw, C., and Bogaardt, M.-J. (2017). Big data in smart farming–a review. Agric. Syst. 153, 69–80. doi: 10.1016/j.agsy.2017.01.023

The What, Why and How of a Digital Preservation DPS

The process by which a Dynamic Purchasing System for Digital Preservation service may (or may not) be adopted

Paul Stokes
Jisc
UK
paul.stokes@jisc.ac.uk
0000-0002-7333-4998

Karen Colbron
Jisc
UK
karen.colbron@jisc.ac.uk
0000-0002-2438-4008

Abstract – Jisc are considering implementing a Dynamic Purchasing system for members to use when procuring a Digital Preservation System. This poster shows what a DPS is and the process we're undergoing to decide if we will provide a DPS. In effect, how we're building a robust business case to provide a DPS for our members.

Keywords – Dynamic Purchasing system, Jisc, Procurement

Conference Topics – SUSTAINABILITY: REAL AND IMAGINED; FROM THEORY TO PRACTICE.

I. Introduction

What is a DPS?

A Dynamic Purchasing system is a procurement framework that simplifies the purchasing process for both buyers and suppliers.

Buyers can quickly procure a digital preservation system using an OJEU compliant process from a pre-qualified set of suppliers.

All the suppliers on the system are verified against a base set of requirements—in this case a base set of requirements for a Digital Preservation System. Jisc recruits the suppliers, ensures compliance with the relevant legislation, and takes care of the required due diligence. Suppliers have reduced cost of sales—the standard due diligence information and base requirements are collected only once and they only need to respond to the requirements that go beyond the base set when bidding—and relatively easy access to Jisc members.

How does it work?

When the time comes to procure a system, buyers run a mini competition using their own overarching set of requirements. They only need to specify requirements that go beyond the base set. Suppliers bid against those requirements. The buyer then selects their chosen supplier and contract directly with them.

For the buyers, a DPS is free to use. It's also considerably cheaper (in terms of resources need to run the procurement) and faster than running an open procurement.

Dynamic?

Unlike traditional frameworks, suppliers can qualify to join a DPS framework at any time in its lifetime. If a supplier is not on the DPS, the buyer Just needs to tell the supplier they want to use the DPS and ask them to complete the application. Assuming the supplier meets the criteria, they can be added very quickly.

II. Making the case

In essence we need to show three things:

1. **Evidence of demand**, both from our members and from the suppliers. A number of complementary channels are being employed including:
    - formal interviews
    - ad hoc discussions

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

- surveys

2. **A market niche** with sufficient numbers which would support such a service.
3. **Economic viability**. This doesn't necessarily mean it should be a profit centre, but there needs to be a good reason for deploying the resources needed to run it.

III. The Poster

The poster will show what a DPS is and the process we have undertaken to make the case including key decision points and information. At the time of writing, the process is still in progress. It is anticipated that it will be completed by the time the poster is published and the final conclusion included.

GIVING ACCESS TO BORN-DIGITAL ARCHIVES AT THE ARCHIVES NATIONALES (FRANCE)

The OeDIPus riddle

Levasseur Emeline
Archives nationales
France
emeline.levasseur@culture.gouv.fr

Falut André
Archives nationales
France
andre.falut@culture.gouv.fr

Fenech Julien
Archives nationales
France
julien.fenech@culture.gouv.fr

Ferrera Matias
Archives nationales
France
matias.ferrera@culture.gouv.fr

Abstract – Born-digital archives, managed in a repository, are accessible thanks to the delivery of Dissemination Information Packages (DIP). While the DIP is a machine-readable format, it is not easily intelligible for the end user. In practice, making DIP truly accessible appears like a riddle. To make it more human-readable, a DIP must undergo some processing outside the repository and be converted in a new form, which should be discussed. These operations call into question the integrity of the archives, which was ensured until their delivery by the system. How, then, can archivists keep data trustworthy? Following recent requests for access, the *Archives nationales* (France) have provided some answers that could be used as a basis for discussion to solve this OeDIPus riddle. Their experience was one of moving from theory to practice, leading to the creation of an in-house proof-of-concept (POC) and tool: OE-DIP (*Objets et Empreintes de DIP*, Objects and Checksums of DIP).

Keywords – access, DIP, integrity, fixity, checksum

Conference Topics – Digital accessibility, inclusion, and diversity; From theory to practice

I. MAKING DIP TRULY ACCESSIBLE: A NEW RIDDLE FOR OEDIPUS

A. How Digital Archives Are Accessed At The Archives Nationales*: Let Oedipus' Journey Begin!*

At the *Archives nationales*, the digital archiving platform, put into service in 2018, is based on the Vitam software (as a back-end). Vitam is an open-source software meeting French and international exchange standards and norms (DEPIP, SEDA, NF Z 42-013) to ensure the ingest, management, preservation and long-term access to digital records of administrative as well as archival value [1]. It is interfaced with the *Archives nationales'* archives management software (as a front-end). The platform can deliver digital archives in a DIP, consisting of a zipped container made up of a directory, without any further structure, containing on one hand the retrieved files, which are renamed in a non-meaningful way, and on the other hand their structural, descriptive, technical and management metadata encoded in a XML file (a "manifest"), in accordance with the SEDA (*Standard d'échange de données pour l'archivage,* Data exchange standard for archiving) [2].

The DIP must be delivered in compliance with the *Code du Patrimoine* access rules [3]. Archives are on principle accessible to all. Therefore, they can be consulted on special workstations in the reading room at the *Archives nationales,* copied on hard drive disks or provided through a secure transfer platform. However, there are many exceptions that prevent archives from being immediately accessible, in order to protect the privacy and safety of individuals or national security. In that case, the reader can ask

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

their producer for an exemption. If it is granted, the reader can be allowed either consultation in the reading room only or full access (consultation and copy).

B. *And Then The Sphinx Asks: "How Can Humans Read And Trust The DIP?"*

The DIP, meant to be interpreted by a machine, is not easily intelligible for end users unfamiliar with digital archiving concepts and data exchange standards. As a result, a transformation is necessary to make the data and their metadata more human-readable. This requires handling of the DIP contents, which can only be carried out outside the repository, after the delivery. However, such an operation is only acceptable if archives can be proven not to be altered in the process. The chain of integrity must not be disrupted [4].

In order to tackle this double challenge - to meet the concrete needs of users while guaranteeing the integrity of the archives they access – the *Archives nationales* have worked on a proof of concept (POC).

II. Oedipus Has Solved The Riddle (Again!)

A. *From DIP To Tree*

The first stage of the POC consists in exporting the archives using ReSIP [5], a tool for processing information packages, usually to prepare SIP (Submission Information Packages), developed by the Vitam program. By interpreting the manifest, ReSIP enables the importation of a DIP and its exportation as a tree of directories and files with their original and meaningful names. A CSV file containing the SEDA descriptive metadata is also exported alongside the tree. This new way of retrieving the archives is easier to understand and seems to meet generic needs, common to all types of users.

B. *OE-DIP: Proving The Archives' Integrity*

The second step of the POC consists in proving that this treatment has not altered the fixity of the accessed archives. To do so, the *Archives nationales* have developed an in-house tool called OE-DIP (*Objets et empreintes de DIP*, Objects and checksums of DIP). OE-DIP performs comparisons between the DIP delivered by the repository on one hand, and the archive tree and its CSV metadata file exported from ReSIP on the other hand. The sequence of instructions first extracts the object hashes from the DIP manifest. It then uses the manifest and the CSV metadata file to locate the corresponding objects in the archive tree, calculates their checksums and compares them with the hashes it has extracted from the manifest. As a result of this comparison, OE-DIP issues an integrity report informing of the outcome of the operation. A positive report acts as a guarantee that the files' fixity was maintained despite the transformation of the DIP. The tool has been tested on various types of digital archives and on large volumes.

III. Solving The Oedipus Riddle, LET'S Not Make A Complex Out Of It

A. *DIP VS Tree: And The Winner Is…*

In the context of this POC, the *Archives nationales* have deliberately delivered digital archives in this double form to researchers: DIP and tree structure. In spite of this double delivery mode, researchers have so far always chosen to access the files in tree form, demonstrating a clear preference for a mode that differs as little as possible from their usual way of browsing. This feedback clearly demonstrates the need to process the DIP and the relevance of delivering the OE-DIP fixity report to users, as a guarantee of trust.

B. *Expected Aftermath*

As of now, ways of improving the OE-DIP tool are under consideration, especially regarding the related issue of metadata integrity. It should be noted that it remains experimental and can only be used in the technical and functional context of the *Archives nationales*. However, if the process meets a need shared by the archives community in France, it could either lead to the development of a hash-calculating feature in the ReSip tool, or to the integration of tree form exports to the Vitam software. The digital archives could then be rendered both as a DIP and directly in a tree form, in order to avoid handling outside the repository, which would enable greater trust and possibly negate the need for the integrity report. In this manner, giving access to digital archives delivered in DIP will no longer represent an Oedipus riddle… nor a complex!

1. REFERENCES

[1] Vitam Program presentation, Programme Vitam. https://www.programmevitam.fr/pages/english/pres_english/

[2] Structuration des Dissemination Information Packages (DIP), Programme Vitam. https://www.programmevitam.fr/ressources/DocCourante/autres/fonctionnel/VITAM_Structuration_des_DIP.pdf
[3] Code du patrimoine: Chapitre 3: Régime de communication (Articles L213-1 à L213-8), Légifance. https://www.legifrance.gouv.fr/codes/section_lc/LEGITEXT000006074236/LEGISCTA000006159942/#LEGISCTA000006159942
[4] Digital Preservation Handbook: Preservation issues, Digital Preservation Coalition. https://www.dpconline.org/handbook/digital-preservation/preservation-issues
[5] ReSIP, Programme Vitam. https://www.programmevitam.fr/pages/ressources/resip/

Establishing an Open-Source Package "Archive"

Euan Cochrane
Yale University Library
USA
euan.cochrane@yale.edu
0000-0001-9772-9743

Rafael Gieschke
University of Freiburg
Germany
rafael.gieschke@rz.uni-freiburg.de
0000-0002-2778-4218

(Extended) Abstract - Many Linux-based operating systems use a package management system that enables users to install a wide range of applications using one or more simple workflows, without needing to find and download the applications from their original publishers. The package management systems also resolve dependencies for users by finding and installing any dependent-applications or "packages" that are needed in order to run the application that the user is trying to install. These workflows greatly improve the experience of working with the operating systems and save a great deal of time for the end-users. When setting up the Emulation as a Service Infrastructure (EaaSI) platform to work with Linux-based operating systems we have encountered a number of issues when working with the operating systems' built-in Package Management Systems (PKMS). The PKMSs usually include a list of servers that host the packages that the PKMS can install for the user. We have found that often those lists are out-of-date and point to servers that no longer exist. In some cases, it is impossible or nearly impossible to find alternative servers that are still actively serving the packages, and where they are available the speed/bandwidth is often much slower/limited for packages for older operating system versions than it was when the operating systems were current. Even when an alternative server can be found that is still actively serving the packages for a legacy operating system, the average end-user can often struggle to understand how to point the PKMS at a custom server as this often requires editing relatively obscure configuration files within the Operating System.

To address these issues the EaaSI program of work and Yale University Library have launched a spin-off project to create a central "archive"[1] of compiled, open-source software packages. The "archive" will host copies of packages for many versions of legacy operating systems and make them freely available to any users, whether working with EaaSI or using the legacy operating systems in other contexts. The archive is hosted directly on a public S3-compatible bucket without an additional frontend server to improve availability, reduce maintainability, allow for practically limitless scaling, and enable users to easily explore and clone its contents (or parts thereof). In addition, the EaaSI software is being updated to enable instance administrators to configure the platform to automatically re-route connections being made to the package servers configured as defaults in legacy Linux-based operating systems and dynamically and seamlessly re-map the connections to the appropriate locations in the new package "archive". This is achieved either by automatically manipulating the DKMS's configuration files or by transparently routing network requests to the originally configured domain to an emulated version of the DKMS's server in a virtual network running in EaaSI.

Unlike Software Heritage [1], the world's preeminent archive of software source code, this project aims to primarily maintain an archive of compiled binaries and not the application source code. While there are some operating systems that use a PKMS that dynamically compiles from source

[1] This term is used very loosly in this context

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

iPRES 2023

code when installing applications, these are rare[2] and we are comfortable with the possibility of overlapping with the work of the software heritage team in this limited area.

There is a wide scope of future work for this project. In spite of being comfortable mildly overlapping on the work of the Software Heritage team, we are also interested in potentially creating reproducible builds of the packages in our "archive" within the EaaSI platform, using source code Sourced from Software Heritage. Doing the compilation from source within the EaaSI platform would provide an audit trail to further strengthen the trustworthiness of the packages provided in our "archive".

Another future extension will be enabling sourcing packages from multiple sources, so that users can use one URL for any version of the same operating system, even if the packages were originally collected from different origin servers by our project. Here, a useful addition could be to seamlessly provide provenance information in S3 for individual packages, i.e., individual files in a directory sourced from multiple different origins.

In the regular use and development of repositories used by PKMSs the specific versions of the packages supplied by the repository are regularly updated and changed. IT is conceivable that users of our archive would need access to a specific version of a package, something that might not be possible if we only included the last version made available in the original repository. Therefore, a further (but much more elaborate) future extension could be to include different versions of the same package, a service offered for Debian at **https://snapshot.debian.org/**.

Further developments could include archiving the repositories of other package managers used in non-operating system contexts, e.g., PKMSs used for acquiring libraries for programming languages, for example., the Python Package Index (PyPI), Node.js' npm, or various Docker registries (e.g., DockerHub, GitHub Container Registry). Here, a particular focus could be on providing time-travel functionality (as offered by snapshot.debian.org) as software projects often do not fully specify their required library versions (e.g., by only specifying a minimum version instead of an exact version) and do not work anymore when newer versions of their required libraries become available at a later time and are selected by the package manager. Here, a bigger focus would probably be on the integration into the system than on duplicating the archive.

In this poster, we will provide a visual overview of the plans outlined in this extended abstract and hope to use it to raise awareness of the "archive" in order to ensure extensive use of it once it is available.

Keywords – emulation, software preservation, open-source

Conference Topics – Sustainability: Real and Imagined; Immersive Information

1. REFERENCES

[1] Software Heritage. Software heritage team. 2023. https://www.softwareheritage.org/ (accessed 9/3/2023)

[2] Many of these PKMS will still serve the automatic build instructions as a special source package as well that would not fall under the scope of Software Heritage.

PROCURING IT SYSTEMS

Thinking about digital preservation from the start

Michael Popham
Digital Preservation Coalition
UK
michael.popham@dpconline.org
0000-0002-6842-4294

Jenny Mitcham
Digital Preservation Coalition
UK
jenny.mitcham@dpconline.org
0000-0003-2884-542X

Paul Wheatley
Digital Preservation Coalition
UK
paul.wheatley@dpconline.org
0000-0002-3839-3298

This poster presents one of the final outputs produced from a collaborative project between the Digital Preservation Coalition and the Nuclear Decommissioning Authority in the UK, "Digital preservation requirements for procuring IT systems". This addition to the Digital Preservation Coalition's Procurement Toolkit proposes the requirements that should be considered when procuring an IT system (for example an EDRMS, DAMS, or GIS) that may ultimately contain at least some records or digital content that needs to be retained beyond the life of the system.

Keywords – Procurement, IT systems, data export

Conference Topics – Sustainability: real and imagined; From Theory to Practice.

I. INTRODUCTION

This poster will summarize the set of six key requirements to consider when procuring any IT system which might contain data of long-term value. It was developed to inform IT procurement practices at the UK's Nuclear Decommissioning Authority (NDA) and was published as an addition to the Digital Preservation Coalition's (DPC) existing Procurement Toolkit [1].

By enabling content extraction in a managed way from IT Systems, organizations can avoid costly barriers to preservation and/or migration of content when an IT system is retired at the end of its life and ensure continued access to valuable information.

The DPC's Procurement Toolkit is intended to provide straightforward advice on how to get the best result out of a procurement process. The bulk of the Toolkit focuses on the procurement of a digital preservation system, but the joint NDA-DPC project "Reliable, Robust and Resilient Digital Infrastructure for Nuclear Decommissioning" (2019-2023) [2] both highlighted the challenges of accessing and secure critical data in legacy systems and aimed to enable the NDA to commission future data and systems with long term resilience from the outset. It seemed helpful to develop some general principles which could be applied to the procurement of *any* IT system which might contain data of lasting value, and it was apparent that advice such as this would benefit many other organizations as well as the NDA.

II. USAGE

The result was a set of six proposed statements of requirement [3] which might usefully be incorporated into the procurement of any IT system – rather than one which was specifically intended for digital preservation. That being said, we encourage users to adapt the language used in these statements to match their organization's particular circumstances, and also to add additional requirements as necessary. These requirements are underpinned by some basic principles, and also accompanied by a statement emphasizing the importance of acceptance testing; these are repeated below.

III. PRINCIPLES

The following principles should be applied to ensure that any content ultimately selected for long-term preservation is managed effectively before transfer to a digital preservation system:

- Appropriate records management policy and procedure should be put in place and fully documented, including clear criteria

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

and related processes for record disposal, retention, and long-term preservation.
- Data and metadata should be structured in a way that makes it straightforward to use and re-use beyond the life of any particular IT system. Open data standards, metadata standards and file formats that facilitate data interoperability are encouraged.
- Robust processes for backing up current data should be applied.

IV. Acceptance Testing

It is essential to verify that a product selected during a procurement process does in fact meet the specified requirements in practice. The inclusion of a content import, extraction, and preservation scenario as part of user acceptance testing is therefore recommended. Testing with a sample set of content and metadata that has been extracted from the IT system can be useful in flagging up any issues before it is too late to make significant changes.

V. Requirements

The six statements of requirement are given below, each with an associated rationale:

1. **The system should use appropriate open data standards to structure and store data.**
   Rationale: Data standards facilitate subsequent data interchange and interoperability without the need for costly and/or complex data migration.
2. **It must be possible to import and store content and associated metadata, if the system is to be populated with existing data.**
   Rationale: If the system to be procured will initially be populated with data from an existing system that it is replacing then it will be necessary to ensure that the data as well as accompanying metadata can be effectively imported and stored.
3. **The system should enable digital content to be selected for disposal or retention/preservation as appropriate:**
   a. Flagging of content by users for action or for specific retention periods.
   b. Selecting content for extraction using search on content and/or metadata.
   Rationale: Not all content held within a system will be of equal value or will need to be kept for the same period of time. Being able to manage retention periods and mark content for deletion or for preservation are important features to help ensure that the right content is managed for the right period of time.
4. **The system must provide a practical mechanism for the extraction of digital content, such as via an API and/or user interface.**
   Rationale: An IT system has a finite lifespan. Suitable export options must be available if the content held within the system has a retention need beyond the life of the system itself.
5. **The system must enable appropriate metadata, structural and contextual information to be extracted along with the digital content.**
   Rationale: Digital content may be of little value without metadata that helps it to be located, understood, and trusted.
6. **The system must allow the extraction of digital content and metadata in formats that will permit its use outside of the system.**
   Rationale: Dependence on an obsolete IT system may hamper or prevent the understanding and use of digital content and metadata.

VI. CONCLUSION

Presenting these requirements as a poster at iPres 2023 will open them up to scrutiny and adoption by a wider audience, and will empower digital preservation practitioners to provide valuable input into procurement processes within their own organizations to ensure that digital preservation requirements are factored into IT system procurement.

1. REFERENCES

[1] https://www.dpconline.org/digipres/implement-digipres/procurement-toolkit
[2] https://www.dpconline.org/digipres/collaborative-projects/nda-project

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

[3] https://www.dpconline.org/digipres/implement-digipres/procurement-toolkit/procurement-toolkit-it-systems

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

EMBEDDING PRESERVABILITY: IFRAMES IN COMPLEX SCHOLARLY PUBLICATIONS

Karen Hanson
Portico
United States
karen.hanson@ithaka.org
https://orcid.org/0000-0002-9354-8328

Jonathan Greenberg
New York University Libraries
United States
jonathan.greenberg@nyu.edu
https://orcid.org/0000-0002-3429-4428

Thib Guicherd-Callin
LOCKSS
United States
thib@stanford.edu
https://orcid.org/0000-0002-6425-4072

Scott Witmer
University of Michigan Library
United States
switmer@umich.edu

Angela T. Spinazzè
ATSPIN consulting
United States
ats@atspin.com

Abstract – As part of a research project, a small team of preservation experts has been embedded within publisher workflows to analyze the challenges associated with preserving complex scholarly publications. As the project reaches the midway point, patterns are emerging regarding preservation-friendly practices that could potentially be incorporated into production processes and platforms to support preservation at scale. One common threat to the preservability of the analyzed publications is the inclusion of web pages that are hosted by a third party (e.g., YouTube videos, ArcGIS visualizations) within the text using *iframes*. The team is exploring methods to improve preservability in such instances while considering the constraints of the project partners and the requirement that preservation services can scale their processes across numerous publications.

Keywords – websites, publishing, publications

Conference Topics – From Theory to Practice; We're All in this Together.

I. BACKGROUND

A new generation of scholarly publications and publishing platforms are leveraging technology to support the integration of complex features, such as embedded streamed audio or video, interactive visualizations, and features for user feedback, into articles and monographs. These dynamic publications create challenges for preservation. The Embedding Preservability for New Forms of Scholarship project [1], which is funded by the Mellon Foundation and led by NYU Libraries, is investigating methods to make these publications more preservable at scale. They are doing this by embedding a team of preservation experts from NYU Libraries, University of Michigan Library, LOCKSS, and Portico into the publishing workflow. For three years, the *embedding team* will shadow the publication production process while engaging with the platforms used by those publishers, interviewing each, and providing feedback as they work on new publications. While an earlier research project developed a framework for understanding the scope of the challenges for preserving complex publications at scale and produced a set of guidelines that publishers could use to improve the preservability of them [2][3], the current project focuses on implementation of the guidelines. Where can the guidelines be integrated into the platform design, user documentation, and publisher workflows? What preservation-friendly practices are most effective, and most likely to be adopted by publishers? As the project approaches the midway point, early patterns indicate some common challenges for preservation and the team is exploring options to manage them. One of these challenges is the use of *iframes*.

II. IFRAMES AND LINK ROT

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

In the publications analyzed, it is common to incorporate complex features from third party platforms using an iframe. An iframe is an HTML tag that allows the author to embed a view of a third-party web page into another web page - in this case, into the text of a publication. Typical examples include 3D ArcGIS visualizations and YouTube videos.

These embedded pages are rarely associated with a persistent URL and are at risk of link rot, where the page moves or is taken offline causing a broken link. This can happen at any time, sometimes before any preservation activity takes place, making these features vulnerable to permanent loss. In addition, the content embedded using an iframe tends to include the most complex and dynamic features in the publication, making them a significant piece of the work, but also sometimes technically difficult to replicate at high fidelity even with the latest web archiving techniques.

III. Need for a Standardized Approach

The embedding team is considering ways to minimize the loss of these resources where they are a core intellectual component of the publication. The team has observed that platforms rarely have sufficient standards or guidance around these integrations to ensure reliable scalable preservation. They are not uniformly managed within or between platforms and often have inadequate or missing captions and/or references. The team's initial suggestions for circumventing loss (e.g., implementing a local web archiving workflow, linking raw data and documentation) were not feasible in the short term for most publishers and platforms, so the team is considering ways to deconstruct these recommendations into smaller steps that require less effort and reduce the impact of this issue.

1) Use of Captions: The convention of adding context and rights information under a figure graphic has not been well adapted for embedding dynamic content. Even though these features are visually displayed, they are often missing captions, and so improving captions for third party content would be a positive step. The team is considering recommendations for appropriate caption content standards, and whether a tool to generate and format a caption, possibly to include machine-readable metadata in the HTML, might be useful. The goal is to include information useful for discovery and understanding of the missing material if the link breaks and leaves a gap in the publication.

2) Alt Text: The practice of adding alt-text to describe non-text features is helpful for accessibility and many publishers are already considering this in their workflows. Alt-text can provide information about embedded content even if it is no longer available. This recommendation would ask that publishers ensure that the iframes' *title* attribute is populated.

3) Data Citation, Archive References: The use of data citation practices to cite the published or archived data source for a visualization using a persistent link is a helpful practice to support preservation. In the case of a GIS visualization, a DOI link might point to a data repository containing the raw data and/or related software. For a non-persistent link to a website, if the resource is compatible with web archiving, this could be a persistent URL to an archived webpage.

4) License Tags: Publishers are used to clearing rights for re-use on graphics embedded in a published analogue work. The nature of the web, however, permits embedding of many resources using iframes without permission. This is convenient, but consequential if the preservation service for that content requires permission from copyright holders or an appropriate license to copy material to the archive. Determining the license of this content at scale requires machine-readable metadata for the object. One proposal is to use the *rel="license"* property in an appropriate HTML tag to designate a license to the embedded content [4]. This would enable the archive to take measured risks when copying the content by automatically reading the tag. It would allow publishers to tag content that should not be copied or be copied but kept in a dark archive until the copyright expires. This property could pair well with URIs from RightsStatements.org, whose purpose is to provide "standardized rights statements that can be used to communicate the copyright and re-use status of digital objects to the public." [5]

IV. Conclusion

The embedding team continues to engage with publishers and the developers of their platforms to explore ways to ensure third-party resources hosted outside of the publisher platform can be preserved.

The second part of the project will determine which of these ideas for managing iframes are practical for publishers and can be easily implemented.

1. REFERENCES

[1] "The Andrew W. Mellon Foundation Awards NYU $502,400 For Libraries Project to Expand Capabilities for Preserving Digital Scholarship," 04-Aug-2021. [Online] Available: https://guides.nyu.edu/blog/The-Andrew-W-Mellon-Foundation-Awards-NYU-502400-For-Libraries-Project-to-Expand-Capabilities-F [Accessed 09-Mar-2023]

[2] J. Greenberg, D. Verhoff, and K. Hanson, "Guidelines for Preserving New Forms of Scholarship," 2021. [Online]. Available: https://doi.org/10.33682/221c-b2xj. [Accessed: 09-Mar-2023].

[3] J. Greenberg, D. Verhoff, and K. Hanson, "Report on Enhancing Services to Preserve New Forms of Scholarship," 2021. [Online]. Available: https://doi.org/10.33682/0dvh-dvr2. [Accessed: 09-Mar-2023].

[4] "CC REL by example." [Online]. Available: https://opensource.creativecommons.org/ccrel-guide/. [Accessed: 09-Mar-2023].

[5] Rightsstatements.org. [Online]. Available: https://rightsstatements.org/en/. [Accessed: 09-Mar-2023].

PLUS ÇA CHANGE...?

Eight years after the end of the 4C project, what next for the Curation Costs Exchange?

Paul Stokes
Jisc
UK
Paul.stokes@jisc.ac.uk
0000-0002-7333-4998

Sarah Middleton
Digital Preservation Coalition (DPC)
UK
sarah.middleton@dpconline.org
0000-0002-7671-403X

Abstract – A poster describing the functionality of the Curations Costs Exchange and the Cost Comparison Tool (outputs of the 4C Project), how they've been used by the community, considerations underway regarding their future, the mechanism used to consult the community and (assuming the consultation and considerations are complete) their ultimate fate.

Keywords – 4C Project, Sustainability, Community consultation, CCEx, CCT

Conference Topics – Sustainability: Real and Imagined; We'er All in This Together.

I. INTRODUCTION

The Curation Costs Exchange (CCEx) is a website [1] created as an output of the 4C project [2].

Understanding the type and magnitude of the costs relating to digital preservation ("digital curation")—both at the time of ingest into a preservation system and ongoing—is one of the thornier problems in digital preservation. The CCEx site addresses this problem. It has two main purposes. It provides information about some fundamental concepts relating to costing digital preservation and it provides a tool that allows registered users to analyse a breakdown of their own costs and make comparisons, either with their own previously entered data or with peer organisations (anonymised).

To these ends the site has 2 main branches, 'Understanding Costs' and 'Comparing Costs'.

1. Understanding Costs

A signposting area that covers the underlying principles and problems associated with costing current and future digital preservation. The four core themes in this section are:

- Basic cost concepts—What are the basics you need to know to get started with costing curation activities?
- Cost models—Are there existing cost models that can help you describe your organisation's activities?
- Cost drivers—How can you be sure that your costs are justifiable? What is that is compelling you to preserve?
- Sustainability planning—What do you need to consider to sustain your organisation's investment in curation?

2. Compare Costs

This area is where the interactive comparisons take place using the Cost Comparison Tool (CCT). Users are prompted for information about their institution, the nature of the digital assets, their preservation practices and resources, and their costs in a number of pre-defined areas.

The information is analysed and normalised in such a way as to allow comparisons, both with peers and with historical information previously entered by the user.

The tool allows users to analyse their costs in a standardised way as well as address the thorny question "how do my costs compare to those of my peers?"

2. THE FUTURE...

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

The CCEx was launched in 2014. Technologies (both web and preservation) have moved on since then and new (additional) concerns have arisen in the area of digital preservation (carbon cost for instance). There are new 'good practice' resources and methodologies available to be signposted.

Currently the interactive cost comparison area of the site (the CCT) is offline. The underlying web framework and modules need to be updated to restore functionality. It could be left in that state. However, there is still cost data in the system which could be updated and/or used by researchers if the appropriate permissions were obtained. It's also worth mentioning that even now, many years after the project concluded, 4C and its outputs are still being referenced in contemporary publications and products.

So, the members of the 4C post project consortium are currently considering the future of the CCEx and the CCT. Broadly speaking there are 4 options:

- Extract the data, back-up the site, and take it off line
- Do nothing—leave the CCEx site operating without the CCT
- Fix the basic problems that prevent the CCT from operating and then do nothing
- Redevelop the CCT to add new functionality

All of these options require the use of resources to a greater or lesser extent. If we're to invest those resources, we need to build a case to do so. With that in mind, we are seeking to gauge the community's interest in the CCEx and CCT by asking the following questions:

- Were you aware of the CCEx/CCT before now?
- Have you used the CCEx/CCT?
- If you've deposited data in the CCT, would you be okay with that data (suitably anonymised) being used as a resources for research purposes?
- Would you like to see the CCT brought back on-line?
- Would you like to see new functions added to the CCT? If so, what?
- Would you be prepared to commit funds to the maintenance/hosting of the CCEx?
- Would you be prepared to commit funds to the maintenance/development of the CCT? If yes, what would your preferred mechanism be?

3. The Poster

At the time of writing this survey and considerations regarding the fate of the CCEx and CCT are ongoing. This poster is intended to describe the evaluation process and ultimate fate of the site and tool.

4. REFERENCES

[1] Curation Costs Exchange—(https://www.curationexchange.org/)

[2] 4C Project—https://www.4cproject.eu/

[3] A more in-depth discussion of the CCEx, CCT and the underlying technology can be found in the 4C deliverable D3.3—Curation Costs Exchange Framework—https://www.4cproject.eu/d3-3-curation-costs-exchange-framework/

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

INTRODUCING TABULA

The University of Minnesota Libraries Digital Preservation System

Carol Kussmann

University of Minnesota Libraries
United States of America
kussmann@umn.edu

Abstract – The University of Minnesota Libraries journey with preserving digital materials has been a long one. After completing an RFP for a preservation system, and then testing that system for multiple years, we decided it was not the system for us. Over 2021 and 2022, we took our requirements along with the lessons we learned from testing, and began to design our own preservation system. Our main goal with this new system is to preserve the unique materials of the Libraries and to be able to provide access to staff that need copies of preservation files for publication or research requests. This poster highlights the development process of Tabula, our digital preservation system.

Keywords – digital preservation, digital preservation system, libraries, implementation

Conference Topics – From Theory to Practice

I. INTRODUCTION

Digital preservation at the University of Minnesota Libraries is managed by the Digital Preservation & Repository Technologies Department (DPRT). DPRT works with a variety of stakeholders across the Libraries and beyond to manage and preserve over 350 TB of materials in all formats.

When we shifted to developing our own digital preservation system, we began by reviewing our requirements and ensuring that we understood the goals and purpose of the preservation system. We focused on five main areas of development utilizing an iterative process: metadata requirements, ingest processes, the hardware/software environment, reporting functionality, and preservation activities.

II. DEVELOPMENT AREAS

A. Metadata Requirements

Existing descriptive metadata schema from multiple sources were studied and crosswalks were developed. The goal was to create a minimal set of descriptive metadata that would assist with the preservation of the materials. With this approach, only two out of 16 various descriptive metadata fields are required, making the system an accessible and effective tool for materials across Library repositories and departments. Administrative and technical metadata requirements were also developed with the goal of long-term preservation activities in mind.

B. Ingest Process

When testing the previous repository, we found that other organizations were performing ingest processes prior to system ingest because it was 'easier' than having the system do it. We wanted to make sure our system did the work for us. Tabula's current ingest process walks through 18 steps that work to add content to the database by assigning ids, associating metadata to the content, verifying that files exist for ingest, performing an anti-virus check, creating/verifying checksums, extracting technical metadata, copying files to multiple storage locations, creating derivatives if needed, and producing a report of the ingest process. Our user interface allows us to build and ingest a SIP as well as check the status of individual SIP steps.

C. Hardware And Software Environment

Encompassing both a web-based graphical user interface and command line menu driven tool, Tabula utilizes the RESTful API Design Methodology [RADM]. At Tabula's core a series of microservices written in the Python scripting language interact with

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

a MySQL database employing a string of RESTFUL APIs built on Spring Boot, a Java-based Framework. Tabula's web interface is built upon Python Flask, a lightweight web application framework, also known as a Web Server Graphical Interface [WSGI].

The modular design of Tabula at the highest level has an external Web Proxy Server, a primary Application Server, secondary Application Servers, an internal proxy server, the RESTful API server, and the MySQL Database Server. The application Servers run Red Hat Enterprise Linux 8 and have 128 GB of RAM and 8 CPUs. The working application space is 10 TB and our "permanent storage" has twin 1 PB of storage and a Tape Library with three LTO9 drives and 100 slots.

D. Database Tables

A series of database tables work together to build the foundation of Tabula. The Element table records information about the 'things' that we want to preserve. Elements are classified as either an Asset or a File. An asset represents one or more files. Both assets and files have their own unique ID numbers.

The Affiliation database documents the names and contact information of organizations, institutions, and repositories for which the elements are related. For example, at the point of ingest we associate materials with the organization from which they came from as well as the access repository in which the materials can be found.

A Metadata table documents descriptive and technical metadata elements available for use within the system. Title, author, description, date of creation, publisher, and geospatial information are some of the descriptive metadata fields. This database also records the technical metadata elements that are captured during the ingest process using DROID.

All actions taken on the Elements are tracked within the Event database. Creating and verifying checksums, performing a virus scan, creating a copy, and moving a file from one location to another, are some of the events that are tracked. We record the type of event, what tool was used, who initiated the event, and the outcome of the event.

Organizing the system with these database tables offers flexibility to develop new tables as needed for additional functionality or needs. We expect to add new tables to assist with our preservation activities.

E. Reporting

We are in the process of developing both internal and external reporting functionality. We want to be able to understand what happens to objects and when, so we document these internal events on SIPs and objects. We also want to be able to answer questions such as: How much content is being preserved? What file formats do we have? When did we receive the materials? We intend to utilize Tableau (a reporting software) to query our databases to produce not only standard reports for our own use but to also create on demand reports in response to questions from our stakeholders about the contents of the repository.

F. Preservation Activities

We are currently building a workbench area where 'problem' objects will be sent to be addressed. The workbench area will be used to address issues discovered upon ingest as well as issues discovered at a later point. It will also be used if a stakeholder needs to review an object. A preservation planning area, an area that will be used to monitor and address preservation concerns, will be developed in the future. We expect to use this area to address objects associated with at risk formats. The preservation planning area will have access to tools to assist with file format identification, migration, and more.

III. Ongoing Work

Developing Tabula continues to be an iterative process. To date, our main focus has been on the environment and ingest process, so we can at minimum preserve our materials. We continue to improve and add functionality to our reporting activities, preservation actions, and the user interface. The end goal is to have a responsive design with UMN branding that will allow users to ingest a set of objects, perform and complete preservation work on one or more of those objects, and allow for non-preservation staff to search, find, and download objects when needed.

METADATA THAT EXCLUDES

A Case Study of the Rock Springs Massacre in Digital Collections

Yingying Han
韩滢莹
University of Illinois Urbana-Champaign
USA
yh17@illinois.edu
0000-0001-6439-9725

Ruohua Han
韩若画
University of Illinois Urbana-Champaign
USA
rhan11@illinois.edu
0000-0001-6321-4194

Karen M. Wickett
University of Illinois Urbana-Champaign
USA
wickett2@illinois.edu
0000-0002-3625-1253

Abstract –This poster presents the preliminary findings of a project that analyzes archival materials about the 1885 Rock Springs Massacre to understand archival silences around the experiences of Chinese people in the United States. We observed that metadata contributes to the exclusion of Chinese people through absent categories, overly narrow subject descriptions, and an emphasis on foreignness and otherness. Future work will identify points of intervention in digital preservation to address these issues.

Keywords – metadata, digital collections, archival silence.

Conference Topics – Digital Accessibility, Inclusion, and Diversity.

I. INTRODUCTION AND BACKGROUND

Trouillot [1] argues that silences enter history at four crucial moments: fact creation, fact assembly (the making of archives), fact retrieval, and retrospective significance. The Rock Springs Massacre occurred in Wyoming in 1885, involving racial violence as white miners attacked Chinese individuals, resulting in the death of 28 Chinese people and displacing hundreds more [2]. The archival silences surrounding this massacre start from fact creation, through witness intimidation and the absence of criminal indictments. The continued lack of evidence and curated collections functions as an "archival amnesty" [3] that enables the United States to evade collective responsibility.

We are library and information scholars working to understand how digital infrastructures can reinforce archival silences. Metadata, an essential element of digital collection infrastructure, shapes the narratives about resources. Giving resources names [4], assigning subject descriptors and designing knowledge organization systems [5] and transcribing historical names into metadata [6] have all been critiqued for their impact on marginalized groups. Studies on reparative and inclusive description have recently gained prominence in archival description and cataloging [7] – [11]. For example, metadata recommendations from the Archives for Black Lives in Philadelphia's Anti-Racist Description Working Group equip professionals to create "ethical, respectful, and accurate description of records created by and about Black people" [11].

As a part of a larger project, we are assessing and reflecting on archival materials about the Rock Springs Massacre collected by four mainstream archival repositories: the Library of Congress, the American Heritage Center (AHC) at the University of Wyoming, the Wyoming State Archives, and the National Archives and Records Administration. We wrote structured reflections for each item about its repository context and metadata, affective responses of ourselves and our imagined audience, and our own motivations and assumptions. Through this process, we examined existing metadata records and contemplated potential interventions within the current descriptive infrastructure. We report initial findings from the critical reflections that focus on metadata and its role in reinforcing archival silences.

II. PRELIMINARY FINDINGS: METADATA AS EXCLUSION

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

A. *Absent Categories*

The issue of uneven representation in navigation categories and topical collections is evident in the Library of Congress's digital collection page[1]. While there is a category for "African American History," there is no equivalent category for Asian American History or other ethnic/racial groups and immigration history. While there is an Asian Division within the "World Cultures & History" topical category, those collections focus on cultural material created in Asia. The observations align with arguments around the invisibility of Asian Americans in archival collections [12], curricula [13], and knowledge organization systems [14].

B. *Overly Narrow Subject Descriptions*

In examining items from the AHC, we observed subject descriptions that primarily focused on the labor dispute aspect of the Rock Springs Massacre. These descriptions highlighted terms such as "Coal Mines and mining," "Chinatowns--United States," and "Immigrants--Wyoming." While these subject headings are not inaccurate, they emphasize a specific narrative surrounding the massacre, centered on industrialization and labor. Alternative subject headings such as "Racism" or "Xenophobia" could provide different perspectives. Future descriptive metadata should incorporate elements embracing counternarratives regarding racism of the massacre, such as narratives about the xenophobia experienced by survivors or victims and presenting a broader perspective from Chinese Americans.

In the Wyoming History Day virtual collection, newspaper clippings from the 1920s mention Chinese men who died in the Rock Springs Massacre. One of these, titled "Lao Chung Dead" includes the metadata Citation "Subject File: Chinese, American Heritage Center, University of Wyoming" and a descriptive note of "The article states that Lao Chung was shot in the back during the 1885 riot and 'carried the bullet to his grave.'". Another clipping tells the story of a man who "hid in a bake oven for three days" during the massacre. We suggest adding subject headings such as "Survivors" or "Survival Narratives".

C. *Othering and Emphasis on Foreignness*

Our examination of digital repositories revealed examples of metadata that emphasized the foreignness and otherness of the Rock Springs Massacre victims. This emphasis, coupled with the focus on labor relations, creates an "archival amnesty" surrounding the event and aligns with arguments around the "perpetual foreigner" status of Asian people in the United States [14]. Five items in the Wyoming History Day virtual collection are categorized under "Photo File: Wyoming-Rock Springs-Foreign Population, American Heritage Center, University of Wyoming", including an engraving from Harper's Weekly, two photographs of the Chinese community's escape route during the massacre, and two pictures of Chinese men returning to China in 1925 and 1926, respectively. Categorizing these items in a "Foreign Population" photo file reflects and reinforces the notion that foreignness is the core theme that ties them together. Especially for the first three items directly related to the massacre, emphasizing the foreignness of the victims in the metadata not only contributes to othering them—it also lacks sensitivity and empathy.

III. Conclusion and Future Work

These initial findings show how metadata structures and terms contribute to the ongoing exclusion of Asian people in the United States through absent categories, overly narrow subject descriptions, and an emphasis on foreignness and otherness. Metadata is not neutral—it shapes our ability to find and understand materials. As we continue this work, we are exploring representation and archival practices around historical materials that foreground radical empathy [15], decolonization, and community. Future work will continue this dialogue between critical examination of digital collections and digital preservation workflows. We aim to identify points of intervention in digital preservation, such as reviewing metadata at ingest, transfer and access and developing processes for community participation around metadata.

REFERENCES

[1] M. Trouillot, *Silencing the Past: Power and the Production of History*. Boston, MA: Beacon Press, 1995.

[2] C. Wu, *"Chink!" A Documentary History of Anti-Chinese Prejudice in America*. New York: World Pub, 1972.

[3] T. Sutherland, "Archival amnesty: In search of Black American transitional and restorative justice," *Journal of Critical Library*

[1] https://www.loc.gov/collections/

and Information Studies, vol. 1, no. 2, 2017. Available: 10.24242/jclis.v1i2.42.
[4] H. A. Olson, *The Power to Name: Locating the Limits of Subject Representation in Libraries*. Dordrecht: Springer Netherlands, 2002.
[5] M. Adler, J. T. Huber, and A. T. Nix, "Stigmatizing disability: Library classifications and the marking and marginalization of books about people with disabilities," *The Library Quarterly*, vol. 87, no. 2, pp. 117–135, Apr. 2017. Available: 10.1086/690734.
[6] R. Han and Y. Han, "Radical empathy in the university archives: Examining archival representations of Chinese students from 1906 to 1920," *Proceedings of the Association for Information Science and Technology*, vol. 58, no. 1, pp. 728–730, Oct. 2021. Available: 10.1002/pra2.543.
[7] L. Hughes-Watkins, "Moving toward a reparative archive: A roadmap for a holistic approach to disrupting homogenous histories in academic repositories and creating inclusive spaces for marginalized voices," *Journal of Contemporary Archival Studies*, vol. 5, no. 1, article 6, 2018. Available: https://elischolar.library.yale.edu/jcas/vol5/iss1/6.
[8] R. L. Frick and M. Proffitt, "Reimagine descriptive workflows: A community-informed agenda for reparative and Inclusive descriptive practice," Dublin, OH: OCLC Research. [Online]. Available: 10.25333/wd4b-bs51. [Accessed: June 29, 2023].
[9] A. Tang, D. Berry, K. Bolding, and R. E. Winston, "Toward culturally competent archival (re)description of marginalized histories," *Library Presentations, Posters, and Videos*, 2018. [Online]. Available: https://digitalcommons.chapman.edu/library_presentations/23. [Accessed: June 29, 2023].
[10] K. Bolding, J. Tai, S. Daniels-Young, and F. Charlton, "Implementing programmatic Anti-Racist (re) description at predominantly White institutions," presented at the ALA 2021 Conference. [online]. Available: https://alair.ala.org/handle/11213/18022. [Accessed: June 29, 2023].
[11] A. A. Antracoli, A. Berdini, K. Bolding, F. Charlton, A. Ferrara, V. Johnson, and K. Rawdon, "Archives for Black lives in Philadelphia: Anti-Racist description resources," 2020. [Online]. Available: https://archivesforblacklives.files.wordpress.com/2020/11/ardr_202010.pdf. [Accessed: May 1, 2022].
[12] M. Caswell, M. Cifor, and M. H. Ramirez, "'To suddenly discover yourself existing': Uncovering the impact of community archives," *The American Archivist*, vol. 79, no. 1, pp. 56–81, 2016. Available: 10.17723/0360-9081.79.1.56.
[13] S. An, "Asian Americans in American history: An AsianCrit perspective on Asian American inclusion in state U.S. history curriculum standards," *Theory & Research in Social Education*, vol. 44, no. 2, pp. 244–276, Apr. 2016. Available: 10.1080/00933104.2016.1170646.
[14] M. Higgins, "Totally invisible: Asian American representation in the Dewey Decimal Classification, 1876-1996," *Knowledge Organization*, vol. 43, no. 8, pp. 609–621, 2016. Available: 10.5771/0943-7444-2016-8-609.
[15] M. Caswell and M. Cifor, "From human rights to feminist ethics: Radical empathy in the archives," *Archivaria*, vol. 81, pp. 23–34, May 2016. Available: https://archivaria.ca/index.php/archivaria/article/view/13557.

A NATIONAL REPOSITORY PLATFORM FOR SHARING THE CHALLENGES OF LONG-TERM DIGITAL CURATION OF RESEARCH DATA

Arif Shaon
Qatar National Library
Qatar
ashaon@qnl.qa

Marcin Werla
Qatar National Library
Qatar
mwerla@qnl.qa

Alwaleed Alkhaja
Qatar National Library
Qatar
aalkhaja@qnl.qa

Abstract – Qatar National Library has launched a research repository named Manara to address the need for curating, preserving, and enabling Open Access to Qatari research output. The repository aims to operate on a consortium-based service model that allows sharing the overall responsibility of curating and preserving a wide variety of research outputs between the Library and key partner institutions, thereby developing a sustainable ecosystem for research outputs in Qatar. The poster presents the work done to establish Manara, including the underlying technical and operational model.

Keywords – Research data, curation, preservation, sustainable model, open access, digital repository

Conference Topics – We're all in this together, Sustainably: Real and Imagined.

I. INTRODUCTION

The research community in Qatar has been producing increasingly large volumes of research output, including both traditional publications and non-traditional scholarly content, such as datasets, over the past decade[1]. This steady growth in Qatar's research landscape is a vital part of the country's effort to transition its economy from a hydrocarbon-based to a knowledge-based one, characterised by innovation, entrepreneurship and excellence in education[2], to achieve a society capable of sustaining its development and providing a high standard of living for its people.

Qatar National Library (QNL)[3] supports Qatar's transition from a natural resource-dependent economy to a diversified and sustainable one through the Library's core values and portfolio of services that include long-term stewardship of and Open Access to Qatari research outputs.

The poster presents the work done by the Library to establish a sustainable research repository for Qatar. The poster highlights the underlying technical and operational model that facilitates sharing the responsibility of long-term digital curation and preservation of Qatari research outputs between the Library and other key stakeholders.

II. THE MAIN CHALLENGE AND STRATEGY

At a very early planning stage, the Library recognised [1] the unique challenges of setting up a national repository service to support various aspects of research output curation, including data and metadata management, handling copyrights and complex licensing models that underpin sharing and publishing different research outputs as well as their long-term digital preservation. These challenges are further magnified by the pervasive lack of awareness regarding the significance of digital preservation, both at a national level and within the broader regional context. This became particularly evident during an online event in 2022 hosted by the

[1] Over 4,000 projects funded that produced over 12,000 publications and 3,000 datasets since 2007. https://app.dimensions.ai/discover/grant?and_facet_funder=grid.507658.9 (Last accessed on 09 March 2023)

[2] Qatar National Vision 2030 - https://www.gco.gov.qa/en/about-qatar/national-vision2030/ (Last accessed on 09 March 2023)

[3] Qatar National Library - https://www.qnl.qa/ (Last accessed on 09 March 2023)

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

Qatar National Library, which brought together repository managers from the region who collectively acknowledged the gap in both comprehension and implementation of digital preservation [2]. Exceptions to this trend include King Abdullah University of Science and Technology (KAUST) in Saudi Arabia for their commendable efforts in this area[4]. Similarly, QNL is the only organisation with an established digital preservation service in Qatar.

In addition, designing a research repository for Qatar requires striking an optimum balance among the diverse needs of the research community, encompassing individual researchers, affiliated institutions, and funding bodies. The complexity of the service infrastructure directly correlates with the wide-ranging nature of the research community. To tackle this challenge, the Library has adopted a strategic approach that involves collaborative partnerships with key stakeholders and the implementation of adaptable and future-proof technical solutions. This approach aims to forge a robust and sustainable service model for a national repository platform for research outputs.

III. Manara – Qatar Research Repository

In November 2022, the Library, in cooperation with Qatar National Research Fund (QNRF)[5], soft-launched a new repository service called Manara[6] (Eng. lighthouse) with the ambition to capture, curate, preserve, and provide (open) access to all research outputs created in Qatar.

The repository provides a range of efficient and user-friendly features designed to support curating and publishing research outputs. It achieves this through a customised version of an institutional instance of Figshare[7], a widely adopted cloud-based repository platform. To ensure the long-term preservation of the deposited content, the repository integrates with the Library's Archivematica[8]-based digital preservation system. This strategic combination of off-the-shelf products, loosely integrated with the existing technical infrastructure, reflects a deliberate design approach. It enables adopting a more suitable repository solution in future while safeguarding against potential vendor lock-in concerns.

A. A Consortium-based service model

The sustainability element of the technical architecture of Manara is effectively reinforced by the service model that leverages a national-level E-resources Consortium that the Library has been coordinating since 2016, ensuring optimum licensing and financial terms for subscription-based access to research and educational content for ten major institutions. Additionally, the consortium actively supports Open Access publishing through read-and-publish agreements. The collaborative nature of this consortium has demonstrated its ability to deliver significant benefits to its members, allowing them to achieve results that would be unattainable if they acted individually.

From an operational viewpoint, Manara is designed as a service which can be operated through a broad partnership, with the Library operating it and any Qatari research-oriented institution using it as its institutional repository. Each participating institution will get a dedicated sub-portal within the repository, necessary training, and content management rights along with digital preservation support from the Library. These dedicated sub-portals can serve as institutional repositories for the partner institutions, facilitating self-archiving (through the Library's digital preservation system) and enabling green Open Access routes for their research outputs.

From a curation standpoint, this consortium-based service model provides a sustainable and cost-effective way of sharing the responsibility of content curation and publishing with the participating institutions. This model places the decision about what to publish within the sub-portals in the hands of the participating institutions and thus fosters a strong sense of responsibility for and ownership of the repository.

[4] KAUST Digital Preservation - https://library.kaust.edu.sa/Digital_Preservation/Overview (Last accessed on 28 June 2023)

[5] QNRF - https://www.qnrf.org/en-us/ (Last accessed on 09 March 2023)

[6] Manara, Qatar Research Repository - https://manara.qnl.qa/ (Last accessed on 09 March 2023)

[7] Figshare - https://figshare.com/ (Last accessed on 09 March 2023)

[8] Archivematica - https://www.archivematica.org/en/ (Last accessed on 09 March 2023)

The Library aims to provide guidance and recommendations regarding content quality, best practices in research data management, copyright and licensing of resources, but participating institutions are ultimately responsible for the content they upload through their designated sub-portals.

IV. FUTURE WORK

Manara is a new service that will initially operate as a pilot. The long-term goal is to develop Manara as a core Library service that is strategically important for the Library's long-term mission and vision, including developing a sustainable ecosystem for managing and disseminating all research outputs in Qatar. The exact long-term sustainability model will be developed in cooperation with participating institutions after the initial pilot period (**by mid-2024**).

1. REFERENCES

[1] Shaon, A., Straube, A., & Chowdhury, K. R. (2018). Setting up a National Research Data Curation Service for Qatar: Challenges and Opportunities. In International Journal of Digital Curation (Vol. 12, Issue 2, pp. 146–156). Edinburgh University Library. https://doi.org/10.2218/ijdc.v12i2.515

[2] Library, Qatar National (2022). Open Repositories in the Middle East and North Africa Meeting, November 2022. Manara - Qatar Research Repository. Collection. https://doi.org/10.57945/manara.qnl.c.6354662.v1

[3] Werla, Marcin; Alkhaja, Alwaleed; Shaon, Arif (2023). Manara - Qatar Research Repository. Manara - Qatar Research Repository. Poster. https://doi.org/10.57945/manara.qnl.23363222.v1

From Redaction to Access

Navigating Challenges to Unlock Houston's LGBTQ Media History

Emily Vinson
University of Houston Libraries
United States
evinson@uh.edu
0000-0002-8902-2454

Bethany Scott
University of Houston Libraries
United States
bscott3@uh.edu
0000-0002-8242-6384

Abstract – *The Gulf Coast LGBT Radio and Television Digitization Project* launched in 2020 to digitize, preserve, describe, and make accessible thousands of hours of Houston's LGBTQ broadcast history. The authors explain the significance of the programs selected for inclusion in the project and describe the steps taken to balance the goal of equitable access to unique materials created by and for a marginalized community, while maintaining adherence to copyright restrictions.

Keywords – LGBTQ Community, Broadcast archives, Audiovisual archives, Digital preservation, Accessibility

Conference Topics – Digital Accessibility, Inclusion, and Diversity

1. Introduction

The University of Houston Libraries (UHL) received the multi-year grant for *The Gulf Coast LGBT Radio and Television Digitization Project* in 2020. This grant, funded by the National Endowment for the Humanities, Division of Preservation and Access, Humanities Collections and Reference Resources Program, supports the digitization, preservation, and online publication of Houston's LGBTQ broadcast history. The selected programs cover a wide range of topics, including politics, activism, health, and LGBTQ identities. This poster provides an overview of the project, its historical context, and the steps taken to ensure access, with a particular focus on our approach to redacting copyrighted content in access copies.

2. Project Overview

Now in its final year, *The Gulf Coast LGBT Radio and Television Digitization and Access Project*, through post-custodial and traditional archives relationships, has reformatted and is in the process of transcribing and describing over 3,500 hours of locally produced radio and television created by and for Houston's LGBTQ community.

The project includes various radio programs produced at Houston's listener-sponsored Pacifica radio station, KPFT, such as "Wilde 'n' Stein," "Lesbian & Gay Voices," and "After Hours." Each program contributed to the representation and empowerment of the LGBTQ community in Houston. The earliest program included in the project, "Wilde 'n' Stein," focused on community outreach, activism, and education about safer sex during the AIDS crisis. "After Hours," a late-night program that aired for thirty years, offered a blend of music, news, activism, entertainment, and an intersectional perspective on the LGBTQ experience.

One television program is included in the project. Broadcast between 1998 and 1999, "TV Montrose" was designed to appeal to all audiences, with a particular focus on LGBTQ life and culture.

A. Project Partners

By combining existing collections with post-custodial collaborations, our goal was to assemble a comprehensive collection of Houston's LGBTQ media history. Donors and project partners had not only collected and housed the source materials, but also contributed to the creation of these unique recordings.

B. Project Workflow

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

As our project launched in May 2020, the challenges caused by the COVID19 pandemic necessitated significant changes to our work plan; however, the broad strokes of our plan and the goals of the project remained the same: to make Houston's LGBTQ media history accessible.

Inventory: Due to the University's hiring freeze, the project PIs took on the task of shifting and inventorying assets from our partner, GCAM. Following safety protocols, we moved analog recordings from GCAM's storage to UHL custody. To expedite the project, we brought tapes home for inventory, transcribing show notes from cassette labels. For digitization, we outsourced to George Blood Audio Video Film, receiving files in various formats with technical metadata and checksums. Samples were assessed for quality assurance.

Redaction: In our planning phase, we obtained permissions from KPFT and the Pacifica Network Archives, but including third-party content in the show episodes posed a challenge for online publication. Licensing agreements held by KPFT did not extend to UHL's digital collections repository, especially concerning commercial music and rebroadcasted news briefs. However, considering the historical significance of the shows, we decided to redact the third-party content to make the recordings available online worldwide. We conducted a fair use analysis and believe that the educational and informational nature of the recordings justifies their online display. The transcripts, descriptive metadata, and transformation of the original works enhance their educational value and promote scholarship.

Though we considered using an AI-based approach to identify music for redaction, we found that most shows included hosts and guests talking over music and original mixes. In order to not remove original show content, we developed a workflow with three student staff members. Over 3,600 files were evaluated for content to be redacted using visual inspection of the waveform and spot-checking the sound. Student employees determined which content required redaction, such as commercial music or news briefs, and which did not, such as the show's theme music. For redacted music, we included a few seconds of the intro and outro with a few seconds of silence in between. These redactions are also noted in the transcripts.

Redacted versions are edited access copies available for online public access. Preservation master files remain unedited, and unedited access copies are accessible in the UHL Special Collections Reading Room or upon request online with restricted access.

Accessibility: Accessibility was a priority, and we aimed to include transcriptions and/or captions for all materials. Machine-generated transcripts alone proved inadequate, leading us to allocate a portion of the budget to professional transcription services. 3PlayMedia was selected to create captions for the television program and transcripts for a portion of the radio programs. For the remaining radio programs, we used an in-house workflow where redacted audio files were submitted to the Otter.ai platform and corrected by student employees. To enhance accuracy, we created vocabulary term lists based on tape label transcripts and shared spreadsheets to track names, places, and acronyms. These terms also inform the descriptive metadata, ensuring consistency throughout the project.

Preservation: We are committed to preserving the media and transcription files included in the project. We have incorporated the project files into the UHL digital preservation program, which follows the OAIS reference model. For each item in the project, we generate a submission information package (SIP) that includes the preservation master file(s), the caption or transcript files, a metadata CSV, and a persistent identifier/permalink. These SIPs are then exported to Archivematica for preservation ingest and storage.

Descriptive Metadata: In collaboration with the UHL metadata unit, we've strived to balance the demands of a large project with the need for descriptive metadata. We include label transcriptions in descriptions when available, and for those without, keywords and host names from transcripts serve as the basis for our metadata records. To ensure inclusive language, we consult subject matter authorities such as Homosaurus and project partners. Metadata records, along with a streaming link to the UHL repository, will be shared with the American Archives of Public Broadcasting for enhanced discoverability.

Publication: UHL's media collections are available online through our instance of Avalon Media System. This system allows us to upload captions for videos

and transcripts for audio recordings in bulk. To support new research methods, such as "collections as data," we will make all transcripts available as plain text files in the UHL Dataverse Repository. This will allow researchers to explore the collection using computational analysis.

In addition to making the recordings accessible to the community that created them, we also hope to increase their reach by connecting with faculty and researchers who may use them for teaching and scholarship. It is only through preservation, long-term access, and reuse of these important collections that we help to diversify the historical record for generations to come.

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

PREMIS IN A PAGE

A beginner's guide to the PREMIS Data Model

Eld Zierau
Royal Danish Library
DK
elzi@kb.dk
0000-0003-3406-3555

Jack O'Sullivan
Preservica Ltd
UK
jack.osullivan@preservica.com
0000-0002-0306-761X

Karin Bredenberg
Kommunalförbundet Sydarkivera
SE
karin.bredenberg@sydarkivera.se
0000-0003-1627-2361

Abstract – This poster from the PREMIS Editorial Committee will provide a high level overview of the PREMIS data model, aimed at providing an introduction to those who are unfamiliar with it.

Keywords – PREMIS, Metadata, Data Model, Preservation Metadata

Conference Topics – From Theory to Practice; We're All in this Together

I. INTRODUCTION

PREMIS (PREservation Metadata: Implementation Strategies) is a de facto international standard for describing Preservation Metadata, i.e. the information that a repository uses to support the digital preservation process. Since its inception, it has been the result of international collaboration between many organizations and institutions, with representation from digital preservation practitioners and from producers and vendors of digital preservation systems.

It is formally defined by the PREMIS Data Dictionary[1], which runs to over two hundred pages. It is also described in an XML Schema and an OWL Ontology. The length and complexity of these three documents presents a high barrier to entry for anyone trying to learn and understand the PREMIS data model.

The core ideas of the PREMIS data model can be presented more succinctly, and independently of either the XML or OWL implementations. This poster will aim to provide an easy-to-follow overview of the most important aspects of the PREMIS data model, without introducing too many technical details of XML or OWL.

II. Description of The Poster

The Data Dictionary itself provides very high level diagrams of the main PREMIS entities, and detailed descriptions of them. It also provides some examples of how specific content can be modelled, however these are also at a detailed level.

This poster is intended to bridge that gap between the very high level, and very detailed. It will consist of a diagrammatic representation of the data model, with brief explanations of each entity, and their relationships to each other, using the diagram in the Data Dictionary as a base for expansion.

It will contain a similar diagram expanding on each of the PREMIS Object types and an explanation of how they are linked together to describe digital content. Again, this will start from the base of the diagram in the data dictionary, but will expand to demonstrate the key, and mandatory attributes of each.

Finally, these concepts will be further illustrated with easy to follow examples, using similar diagrams for consistency.

III. Intended Audience

This poster is mainly targeting 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) who are not familiar with PREMIS and are seeking a high level overview.

1. REFERENCES

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

iPRES 2023

[1] PREMIS Editorial Committee. 2015. PREMIS Data Dictionary for Preservation Metadata. Accessed 2023 located at **http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf**, Web archived: archive.org,, archive time: 2017-02-10 06:23:29 UTC archived URL: **http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf**.

PREMIS Implementation Fair Workshop

Exchanging Experiences using PREMIS

Micky Lindlar
TIB Leibniz Information Centre for Science and Technology
Germany
michelle.lindlar@tib.eu
0000-0003-3709-5608

Karin Bredenberg
Kommunalförbundet Sydarkivera
Sweden
karin.bredenberg@sydarkivera.se
0000-0003-1627-2361

Sarah Romkey
Artefactual Systems
Canada
sromkey@artefactual.com
0000-0003-3833-7648

Marjolein Steeman
Netherlands Institute for Sound&Vision
the Netherlands
msteeman@beeldengeluid.nl
0000-0002-1506-1581

Abstract – This workshop functions as a meeting ground for PREMIS practitioners, researchers, and curious onlookers. As opposed to PREMIS tutorials, which focus on introducing the data model and data dictionary, the PREMIS Implementation Fair focuses on work currently undertaken by the Editorial Committee as well as on real-world implementations, projects and open questions that the digital preservation community has around preservation metadata.

Keywords – PREMIS, Preservation, Metadata, Preservation Metadata

Conference Topics – From Theory to Practice; We're All in this Together.

I. Introduction

The PREMIS Data Dictionary for Preservation Metadata [1] is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. It is maintained by the PREMIS Editorial Committee and the PREMIS Maintenance Activity is managed by the Library of Congress [2].

The PREMIS Implementation Fair Workshop is one of a series of events organized by the PREMIS Editorial Committee and in conjunction with iPRES conferences. While the PREMIS Tutorials focus on giving in-depth understanding of what PREMIS is and how it can be used, the PREMIS Implementation Fairs bring together digital preservation practitioners and researchers who are already using PREMIS or are planning to do so and want to share their thoughts and projects or pose a question to an audience of PREMIS implementers.

With the last iPRES PREMIS Implementation Fair having been held as a "pop-up event" during iPRES 2018, much has happened in the digital preservation field - and in PREMIS - since. At iPRES 2023, the workshop will give implementers, and potential implementers, of PREMIS an opportunity to discuss topics of common interest and find out about latest developments. The event will pick up on three topics which the digital preservation community has voiced particular interest in: PREMIS and emulation, PREMIS in end-to-end systems, and PREMIS implementations for rights.

II. Form of the Tutorial

The Implementation Fair will be held as a 90-minute on-premise workshop.The workshop is planned in two parts: set presentations, which focus on different themes for which potential speakers have already been identified as well as an "open" session, for which short presentations / questions will be solicited from workshop participants as well as the wider digital preservation community before the event.

Reasoning behind these two parts is balancing open workshop participation and flexibility with mitigating the risk of little participatory input.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

III. Summary of the Workshop

The workshop starts with a brief introduction to the Implementation Fair format and the plan for the next 90 minutes. The Implementation Fair will focus on real world implementations and existing questions that practitioners have been facing in the integration of preservation metadata into their workflows and practices.
The first three showcasepresentations will focus on one or more of three strands. These strands will be briefly introduced in the beginning of the workshop:

- PREMIS in emulation
  While PREMIS has the means to capture preservation metadata for emulation, few real-world implementations are known. How is PREMIS currently used for emulation? What are barriers to not use it?
- PREMIS in end-to-end systems
  A lot of users use PREMIS without knowing it. This is especially true in big graphical user interface heavy end-to-end systems. What does a PREMIS implementation look like under the hood? Where does it work or where does it currently fall short?
- Implementations of using PREMIS for rights
  Do institutions currently use PREMIS to capture rights information about digital objects? If so, how? If not, why not? Are there cases where users are not sure if PREMIS is a fit for rights or not?

The introduction is followed by three presentations who each touch on at least two of the three strands:

- An Archivematica community user will showcase the PREMIS in METS implementation used in Archivematica.
- Micky Lindlar of TIB will showcase how PREMIS is used in TIB's Rosetta-based digital archive and give insight into current limitations in capturing rights
- An Preservica community user from Yale will showcase how PREMIS is used in Preservica and the status-quo of PREMIS for emulation

While these three speakers are pre-identified, the Implementation Fair is planned as a very interactive workshop which should allow for widest possible community participation. Upon acceptance of the workshop proposal, the organizers will solicit contributions to the workshop via an open community call. The organizers will try to accommodate all community proposals in the workshop session. If it so happens that more proposals exist than time allows for, the organizers will ensure that the selection of contributions represent the diversity of the digital preservation community in regard to institutional types, sizes and maturity of the digital preservation implementations ranging from "planned" to "in production". In the case that not enough contributions are submitted, the organizers will make strong use of interactive methods such as Mentimeter polls and open discussion sections during the workshop to allow different participation methods for the audience.

IV. Agenda

1. Introduction to the Workshop
2. Showcase presentations
3. Open Participation
4. Summary and Conclusion by Workshop facilitators

V. Intended Audience

This workshop is mainly targeting 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 in digital preservation metadata and preservation risk assessment with the aim of providing a platform to discuss topics of common interest and find out about latest developments.

1. REFERENCES

[1] PREMIS Editorial Committee. 2015. PREMIS Data Dictionary for Preservation Metadata. Accessed 2023 located at **http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf**, Web archived: archive.org,, archive time: 2017-02-10 06:23:29 UTC archived URL: **http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf**.

[2] PREMIS website. Accessed 2023. Located at **http://www.loc.gov/standards/premis/index.html**.

SERVER-SIDE WEB ARCHIVING USING REPROZIP-WEB

Katherine Boss
New York University
USA
katherine.boss@nyu.edu
0000-0003-2148-8386

Ilya Kreymer
Webrecorder Software
USA
ilya@webrecorder.net

Vicky Rampin
New York University
USA
vicky.rampin@nyu.edu
0000-0003-4298-168X

Rémi Rampin
New York University
USA
remi.rampin@nyu.edu
0000-0002-0524-2282

Abstract – Current client-side, or "static," web archiving crawlers have been tremendously successful in capturing and archiving millions of pages of the internet. Unfortunately, over the decades the web has evolved beyond the reach of many of these crawlers, and today's static crawlers fail to capture the look, feel, and functionality of a significant amount of interactive web content, including maps, visualizations, database-reliant projects and social media feeds. Archiving these dynamic websites requires a different approach, including a server-side web archiving option.

ReproZip-Web is an open source, grant-funded [1] web-archiving tool that can address this need. It builds on the high-fidelity crawling tools of Webrecorder by also encapsulating a dynamic web server software and its dependencies. The output is a self-contained, isolated, and preservation-ready bundle, an .rpz file, with all the information needed to replay a website, including the source code, the computational environment (e.g., the operating system, software libraries) and the files used by the app (e.g. data, static files). Its lightweight nature makes it ideal for distribution and preservation.

This interactive workshop will be particularly useful for web archivists, digital archivists, digital humanities scholars and others seeking to archive and preserve complex web projects. Attendees, who should be familiar with the command line interface, will practice packing and tracing a web application and recording the front-end of the site using ReproZip-Web. They will then be able to test replaying the site from the newly created and preservable .rpz file.

.

Keywords – Dynamic web archiving, Server-side web archiving, ReproZip-Web, Webrecorder

Conference Topics – We're All in This Together; Immersive Information

1. REFERENCES

[1] Institute of Museum and Library Services, "Preserving the Dynamic Web: Building a Production-level Tool to Save Data Journalism and Interactive Scholarship," *NLG-L Recipient, LG-250049-OLS-21*, August, 2021. http://www.imls.gov/grants/awarded/lg-250049-ols-21 (accessed Mar. 10, 2023).

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

iPRES 2023

PRESERVING EMAIL ATTACHMENTS WITH ATTACHMENT CONVERTER

Matt Teichman
Digital Library Development Center
University of Chicago Library
USA
teichman@uchicago.edu
https://orcid.org/0000-0002-9637-6613

Ashley Gosselar
Hanna Holborn Gray Special Collections Research Center
University of Chicago Library
USA
agosselar@uchicago.edu
https://orcid.org/0009-0000-4172-7618

Obi Obetta
Digital Library Development Center
University of Chicago Library
USA
obettasteph@uchicago.edu

Nishchay Karle
Digital Library Development Center
University of Chicago Library
USA
nishchaykarle@uchicago.edu

Keywords – email, MIME, attachment, digital preservation, digital formats

Conference Topics – Sustainability: Real and Imagined ; From Theory to Practice

I. WORKSHOP CONTENT

Email increasingly forms a core portion of many archives, and while tools are available for basic preservation of email messages, less consideration has been given to archival preservation and control of email attachments. It is essential to future research that these important records are preserved in the context of their associated messages.

With the generous support of an "Email Archives: Building Capacity and Community" grant from the University of Illinois, the University of Chicago Digital Library Development Center has created a free and open-source tool called Attachment Converter. Attachment Converter utilizes file format conversion utilities already installed on a user's system to batch-convert common email attachments to formats recommended for archival preservation and access, retaining the connection between the migrated attachments and their associated emails.

In this workshop, iPres attendees will learn the basic backend anatomy of an email, how to convert email from PST to MBOX, and inspect and analyze the contents for common attachments. We will demonstrate how to migrate attachments to preservation formats using Attachment Converter and discuss how archivists without UNIX/Linux experience can collaborate with their institution's IT staff to implement the tool.

II. WORKSHOP AGENDA

The workshop will have two parts. First, we will explore the basic technical components of email and share techniques for working with them. The full email specification is too complicated for a 90-minute workshop, but we will provide everything that an archivist needs to know for the purpose of working with email attachments. The goal is for all attendees to walk away with the ability to convert email from Outlook PST to MBOX format, open the MBOX in a text editor, and inspect the raw contents of the mailbox, regardless of their technical background. Second, we will teach attendees how to set up and use Attachment Converter.

The workshop is designed to be interactive and attendees are encouraged to follow along on their computers using a sample mailbox that will be provided to them. The sample emails will contain attachments in file formats archivists may commonly wish to preserve: JPEG, GIF, PCX, PDF, DOC, DOCX, RTF, XLS, and XLSX. We will show participants how to

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

open a mailbox in a text editor and identify where the attachments are in an email, how to guess the file format of an attachment, and also explain why that has to be a guess. We will then demonstrate how to convert all the attachments in a single email to archivally stable formats using Attachment Converter. Following that, we will show how to batch-convert attachments in an entire mailbox.

Finally, we will explain how to set Attachment Converter up to use a utility of the user's choice to convert attachments in their emails. This process requires more technical expertise---so we won't go into full detail---but we will explain that feature at a high level and stress that any archivist working somewhere with an IT staff that knows UNIX/Linux system administration should be able to make use of this feature.

III. Learning Objectives

At the end of this workshop participants will be able to:

- Understand the structure of an email;
- Convert email from Outlook PST to MBOX format;
- Open and analyze MBOX email in a text editor;
- Install and run Attachment Converter on a sample MBOX (provided);
- Communicate with IT staff about configuring Attachment Converter for local needs.

IV. Workshop Requirements

This will be a 90-minute virtual workshop. Students will need the following:

- A computer (Mac or Windows) and internet connection, and the ability to download files from Box;
- A text editor (Notepad or Text Edit);
- Optional advanced: If students want to follow along, they will need admin rights on their computer, and will need to install Homebrew (Mac) or WSL Ubuntu (Windows). Installation guides will be provided beforehand via Box.

Attachment Converter will be demonstrated in a UNIX environment. Knowledge of how to use UNIX is not required for this workshop, but basic knowledge of command line would be useful for understanding the demonstration.

STORAGE AND STANDARDS:

Shaping Version 4 of the DP Storage Criteria

Andrea Goethals
National Library of New Zealand
New Zealand
Andrea.Goethals@dia.govt.nz
0000-0002-5254-9818

Cynthia Wu
National Library of New Zealand
New Zealand
Cynthia.Wu@dia.govt.nz
0000-0002-9318-275X

Eld Zierau
Royal Danish Library
Denmark
elzi@kb.dk
0000-0003-3406-3555

Sibyl Schaefer
University of California San Diego
USA
sschaefer@ucsd.edu
0000-0002-7292-9287

Nancy McGovern
Global Archivist LLC
USA
mcgovern60@gmail.com
0000-0002-7733-1516

Abstract – The Digital Preservation Storage Criteria ("Criteria") are designed to provide community guidance for organizations that either use or provide digital preservation storage. Currently on its fourth iteration, the Criteria have been mapped to relevant information security and digital preservation standards such as ISO 14721 (OAIS) and ISO 16363 (Audit and certification of trustworthy digital repositories). This standards mapping exercise has led to many changes in the Criteria. The authors propose a workshop to get feedback from participants on key elements of the draft fourth version of the Digital Preservation Storage Criteria ("Criteria").

This workshop will provide participants with an introduction to the Criteria and then will lead the participants in group exercises designed to test the assumptions and work behind draft version 4 of the Criteria. The expected outcomes are a deeper understanding of considerations for digital preservation storage by the participants, and feedback to the Criteria working group on draft version 4 before it is published after the conference.

Keywords – preservation storage, standards

Conference Topics – We're all in this together; From theory to practice.

I. INTRODUCTION

The Digital Preservation Storage Criteria [1] grew out of an iPRES 2015 community discussion on the various and evolving approaches to digital preservation storage. The discussion participants identified a gap in guidance for organizations that either use or provide storage for digital material that must be preserved long-term. An international working group of volunteers formed to develop what came to be the Digital Preservation Storage Criteria ("Criteria") and accompanying Usage Guide. From 2016-2019 the working group produced three iterative versions of the Criteria, each time gathering feedback at digital preservation conferences to improve the next version.

From 2020-2022 the working group mapped the Criteria to relevant information technology and digital preservation standards, such as ISO 14721 [2] and ISO 16363 [3]. The standards mapping process has resulted in many changes to the Criteria that will become Version 4. A review of these standards revealed where there were missing criteria, insufficient criteria definitions, and even missing categories of criteria. This workshop will give participants a chance to review and give feedback on draft Version 4 so that this feedback can be incorporated into the published version after the conference.

The target audience for this workshop are organizations that require or provide storage for digital content with long-term preservation needs. Because an introduction to the project will be provided at the beginning of the workshop, no prior knowledge of the Criteria is required to attend.

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

iPRES 2023

II. Workshop Description

A. Format and Length

The authors propose either one or two ninety-minute blocks. If two blocks are provided, the first block will be an overview of the Criteria - how it came about, what it is intended for, how it is being used within organizations, and recent developments. After the break, the second block will be a series of group exercises to get feedback on draft Version 4 of the Criteria. If only one block is provided, the format will be the same, but the overview of the Criteria will be shortened to a thirty-minute introduction followed by an hour group exercise.

Props will be created for the group exercises, for example, printing the criteria and categories on cards that can be sorted and arranged. Participants will be divided into three to four groups (depending on the number of participants) to work through the exercises.

After the group exercises, the groups will come together to report back on results and discuss changes suggested by the groups. An open discussion will follow so that participants have an opportunity to give their overall feedback on draft Version 4. The feedback from the group exercises and open discussion will be incorporated into the published version after the conference.

B. Group Exercises

The Criteria are composed of 74 criteria, each with a name and definition. Each criterion is mapped to a single category, which also has a name and definition. Seven information security and digital preservation standards were reviewed, and excerpts of the standards were mapped to the relevant criteria. Group exercises have been designed to get feedback on the following questions.

1. Exercise 1 - Are the criteria well-named and well-defined?

Participants will be asked to match criteria names to criteria definitions to see if there are any difficulties in understanding the text used to name and define the criteria. Participants will also be asked if the definitions could be improved.

2. Exercise 2 - Are the categories well-named and well-defined?

Participants will be asked to match category names to category definitions to see if there are any difficulties understanding the text used to name and define the category. Participants will also be asked if the definitions could be improved.

3. Exercise 3 - Are the criteria placed in the appropriate categories?

Participants will be asked to match the criteria to the closest logical category to see where there is consensus and differing opinions. Where there are differing opinions, we will dive into the reasons for this.

4. Exercise 4 - Do the standards map well to the relevant criteria?

For key standards or excerpts, for example, key text from ISO 16363 [3], participants will be asked to map the excerpt to the relevant criteria. The purpose will be to see if participants mapped the excerpts to the same criteria as the working group in creating draft Version 4.

III. References

[1] Schaefer, S. K., McGovern, N. Y., Zierau, E. M. O., Goethals, A. L., & Wu, C. C. M. (2022). Deciding how to decide: Using the Digital Preservation Storage Criteria. IFLA Journal, 48(2), 318–331.

[2] *Space data and information transfer systems – Open archival information system (OAIS) – Reference model*, ISO 14721, 2012.

[3] *Space data and information transfer systems – Audit and certification of trustworthy digital repositories*, ISO 16363, 2012.

THIS IS FEDORA 6.X

Understanding the Oxford Common File Layout, Intro to Migration Tools and Understanding Community Developed Integrations

Arran Griffith
LYRASIS
Canada
arran.griffith@lyrasis.org

Dan Field
LYRASIS
Wales
dan.field@lyrasis.org

Fedora 6.x is the newest, most modern version of the software, representing a significant change in the preservation standards and backend infrastructure from previous versions. This modernization of the software provides users a more robust preservation platform, while giving the community back the data transparency they appreciated from Fedora 3. This workshop will provide participants with the ability to work directly with the newest version of Fedora through hands-on exercises as well as learning about the Oxford Common File Layout (OCFL) and it's role in digital preservation within a Fedora ecosystem. We will complete a sample migration using the Migration Toolkit, a series of instructional modules created from an IMLS grant-funded project, and gain experience working with the migration utility and validator tools. Lastly we will explore several community-developed integrations that allow for additional functionality and visibility into the contents within a Fedora repository.

This workshop is intended to provide participants an opportunity to work directly with the software, understand it's preservation features and become familiar with the recently developed tools for migrating from Fedora 3.x to 6.x.

Keywords: Fedora, Repository, Open Source, Migrations

Conference Topics –From Theory to Practice, Immersive Information

I. INTRODUCTION

In July, 2021, the long-awaited Fedora 6.0 was released. This workshop will provide an overview of the software itself, a look at our roadmap and path to release, as well as dive into some important new features that helped return Fedora to it's digital preservation roots. We will showcase and demonstrate the newly available migration tooling and documentation as we work through a hands-on migration. Lastly we will demonstrate how to integrate Fedora with your ecosystem via the Camel Toolbox.

The workshop will include several hands-on portions that will allow attendees to exercise Fedora features, while learning about their purpose and function. These features are accessible via a built-in web interface, so no command line experience is required.

II. HANDS-ON BREAKDOWN

We propose to break the hands-on portion down in to the following segments for easier comprehension:

Section 1: Fedora 6 Technical Overview & Resources Management

- Highlight and test new features of Fedora 6.x and understanding how to work with resources within the Fedora platform.

- Outline and explain OCFL's role in digital preservation within Fedora and how users

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

can view and interact with OCFL files on disk.

Section 2: Migration

- Participants will engage in a migration of a sample data set from Fedora 3.x to Fedora 6.3 using the migration tooling.

- Participants will be given an opportunity to work through components of the Migration Toolkit guided by the facilitator.

Section 3: Fedora and Community Integrations

- Understanding community-developed integrations like the Camel Toolbox and others, demonstrating how to integrate Fedora with your ecosystem using it.

This is a technical workshop for those with command line experience. While no explicit Fedora experience is required, a general understanding of the role, components and functionalities of a repository would be beneficial. Attendees who wish to participate in the hands-on sections will need to access an online sandbox via a URL which will be provided ahead of the workshop. Participants will be required to bring their own laptop for participation - tablets and handheld devices will not be supported.

III. Learning Outcomes

1. By the end of this workshop, participants will be able to comfortably manage resources within Fedora and understand how Fedora provides a digital preservation solution using the features available in the newest release - namely how OCFL provides a robust, transparent and long-lasting solution.
2. Participants will also be familiar with the Fedora migration tool suite and how to use it to execute a Fedora 3 - 6 migration as well as how to integrate extensions into their Fedora environments.

REFERENCES

[1] Manuscript Templates for Conference Proceedings, IEEE. http://www.ieee.org/conferences_events/conferences/publishing/templates.html

Continuous Improvement Tools for Developing Capacity and Skills

A Tutorial

Sharon McMeekin
Digital Preservation Coalition
Scotland
sharon.mcmeekin@dpconline.org
0000-0002-1842-611X

Jenny Mitcham
Digital Preservation Coalition
United Kingdom
jenny.mitcham@dpconline.org
0000-0003-2884-542X

Amy Currie
Digital Preservation Coalition
Scotland
amy.currie@dpconline.org
0000-0001-9099-8457

Abstract – The ability to apply a carefully considered and well implemented approach to continuous improvement of digital preservation capabilities can greatly benefit practitioners when looking to set and achieve objectives. This tutorial aims to provide attendees with the skills and tools to develop and implement a methodology for continuous improvement at their organization using resources developed by the Digital Preservation Coalition. This will include assessing maturity with the Rapid Assessment Model and auditing skills with the Competency Framework and Audit Toolkit.

Keywords – Maturity modelling, skills, good practice, continuous development, benchmarking

Conference Topics – Sustainability: Real and Imagined, Digital Accessibility, Inclusion, and Diversity, From Theory to Practice.

I. Introduction

Digital preservation cannot be a static activity. Ensuring the longevity of digital content requires proactive management and maintenance of the organizational and technological infrastructures we deploy. But how best to structure this management and maintenance to ensure its success?

Since the early days of digital preservation, the community of practice has sought ways to benchmark an organization's capabilities. An audit and certification approach was the original method championed, however, in recent years the more flexible approach of maturity modelling has started to gain popularity. A maturity model provides a framework for assessing the level of capability of an organization across defined areas relating to policy, processes, procedures, and infrastructure. Maturity models allow an organization to understand their current capabilities, set future targets, and plan for developments to meet those targets.

As part of their member support activities, the Digital Preservation Coalition (DPC) has created a number of resources to facilitate the continued development of digital preservation capabilities within an organization. These include the DPC Rapid Assessment Model[1] (DPC RAM) and the DPC Competency Framework[2]. These two resources are the focus of the proposed tutorial.

II. Summary of the Tutorial

This aim of this tutorial is to empower practitioners by providing them with the tools and skills required to plan, advocate for, and assess their progress with developing digital preservation capabilities within their organization.

It will begin by providing them with a solid understanding of the importance and benefits of a continuous improvement approach to benchmarking their digital preservation capabilities. Following this, attendees will be introduced to and led through two practical exercises:

[1] https://www.dpconline.org/digipres/implement-digipres/dpc-ram

[2] https://www.dpconline.org/digipres/train-your-staff/dp-competency

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

1. Using DPC RAM to assess an organization's capabilities with reference to policy, processes, procedures, and infrastructure.
2. Carrying out either an individual or organizational skills audit using the DPC's Competency Framework and Audit Toolkit.

As well as practical exercises, attendees will be encouraged to engage with live polling to allow benchmarking of digital preservation maturity of the organizations represented within the tutorial cohort.

The tutorial will finish with an overview of other DPC resources that can help practitioners with planning and advocating for their digital preservation activities.

III. Content Outline

The following is a draft outline of the tutorial content, including proposed timings:

1. Intro. to Continuous Improvement (c. 30mins)
    a. What is continuous improvement?
    b. Benefits of a continuous improvement
    c. Introduction to continuous improvement tools from the DPC
2. Focus on DPC RAM (c. 60mins)
    a. Introduction to DPC RAM
    b. Exercise: completing a DPC RAM assessment
3. Break
4. Focus on the DPC Skills Framework (c. 60mins)
    a. Introduction to the DPC Competency Framework and Audit Toolkit
    b. Exercise: completing a personal or organizational skills audit
5. Feedback and Wrap-Up (c. 30mins)
    a. Overview of DPC resources to support continuous improvement
    b. Tutorial feedback

IV. Intended audience

This tutorial will benefit individuals and organizations from across many sectors who wish to assess their current digital preservation capabilities and plan for future developments. It will also benefit researchers wishing to incorporate an understanding of these processes into their work, and educators who hope to expand or enhance their curricula on the topics covered.

V. Learning Outcomes

Tutorial attendees will be able to:

3. Explain the importance of continuous improvement.
4. Plan their approach to continuous improvement.
5. Complete a DPC RAM assessment.
6. Describe the skills required for preservation.
7. Undertake a skills audit of digital preservation staff at their organization.

VI. Short Biographies of Organizers

Sharon McMeekin is Head of Workforce Development at the DPC, which includes acting as managing editor of the 'Digital Preservation Handbook', and lead author and project manager of the Novice to Know-How training resources. Sharon is an archivist and experienced practitioner, has contributed to international training and development projects, and is a frequent guest lecturer for information management courses.

Jenny Mitcham is Head of Good Practice and Standards at the DPC where she engages in a range of projects to develop good practice resources for digital preservation. This has included a project working with the UK Nuclear Decommissioning Authority, during which she co-created DPC RAM. Jenny has worked in digital preservation for nearly two decades, having previously held roles at the Archaeology Data Service and the University of York.

Amy Currie is Training and Grants Manager at the DPC, where she works on the development of digital preservation training and skills projects and manages the Career Development Fund. She completed her PhD at the University of Glasgow in 2021, where she previously worked as a teaching assistant and co-convenor in the Information Studies department.

UNDERSTANDING AND IMPLEMENTING METS

A tutorial focused on METS 2

Karin Bredenberg
Sydarkivera
Sweden
karin.bredenberg@sydarkivera.se
0000-0003-1627-2361

Aaron Elkiss
HathiTrust
USA
aelkiss@hathitrust.org
0000-0002-2904-9559

Juha Lehtonen
CSC - IT Center for Science
Finland
juha.lehtonen@csc.fi
0000-0002-9916-5731

Abstract – This half day tutorial will provide participants with an introduction to the Metadata Encoding and Transmission Standard (METS) starting in METS version 1 and the METS Primer [1], but focusing on METS version 2. It will give a basic overview of METS and explore different models of implementation. The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library as well as digital archives, expressed using the XML schema language of the World Wide Web Consortium. It is maintained by the METS Editorial Board, and the METS Maintenance Activity is managed by the Library of Congress [2].

Keywords – Metadata and information strategies and workflows; Infrastructure, systems, and tools; Case studies, best practices and novel challenges; Training and education for a new version

Conference Topics – We're All in this Together; From Theory to Practice

I. INTRODUCTION

METS, the Metadata Encoding and Transmission Standard, provides a key piece of infrastructure for digital transfer as well as digital preservation activities, playing a vital role in enabling the effective management, discovery, and re-usability of digital information. METS continues to be widely used in digital preservation to describe the contents and structure of digital objects [3] and to provide descriptive, administrative, and structural information about these objects. By working in conjunction with other standards, METS gives information regarding documents preservation activity, identifies technical features, and aids in verifying the authenticity of digital objects. METS contains a set of metadata elements recommended for use in all transfer as well as archiving situations regardless of the type of materials being transferred or archived, the type of institution, and the transfer strategies employed.

II. SUMMARY OF TUTORIAL

METS can be used to describe objects in the role of Submission Information Package (SIP), Archival Information Package (AIP), or Dissemination Information Package (DIP) within the Open Archival Information System (OAIS) Reference Model. METS version 2 [4, 5] simplifies the schema, makes it more consistent, and removes reliance on the outdated XLink standard. It aims to retain a clear path for migration from METS 1 for most use cases.

This tutorial introduces METS with a focus on version 2 and its elements. The tutorial will introduce the elements of METS and the changes between version 1 and 2. It will include methods and examples for migrating from version 1 to 2. In addition, it will present examples of METS metadata and a discussion of implementation considerations, particularly using METS in combination with the other XML metadata standards such as the "Preservation Metadata: Implementation Strategies" (PREMIS) [6] standard. It will include examples of transformations from existing METS 1 objects to the new METS 2.

The tutorial aims to develop and spread awareness and knowledge about metadata that supports transfer and long-term preservation of digital objects, regardless of the version of METS in use.

III. CONTENT OUTLINE

The draft outline for the tutorial is below.

A. Introduction to METS

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

1) Background (brief history and rationale)

2) Status of METS

3) Benefits of implementing METS 2

B. METS in detail with a focus on METS 2

1) Core elements and a simple example

C. Implementation

1) METS 1 to METS 2

2) The case of using PREMIS and other metadata standards in METS

3) Support and the METS community

4) Conformance

D. Next steps and wrap up

1) Round table discussion for institutional plans

IV. INTENDED AUDIENCE

This tutorial will benefit individuals and institutions interested in learning about METS but who have limited experience in implementation as well as those interested in potential migration paths to METS 2. The tutorial will cover implementing METS for transfer as well as for the long-term management and preservation of digital information. The potential audience includes cultural heritage operators, researchers and technology developers, professional educators, and others involved in management and preservation of digital resources.

V. EXPECTED LEARNING OUTCOMES

A. Participants will understand:

1) What METS is and why it exists;

2) The benefits of implementing METS;

3) The differences between the two versions of METS;

4) The nature of the existing METS community;

5) The critical role METS plays in the digital preservation community for transferring digital objects.

B. In addition, participants will get insight into:

1) How METS may be used with PREMIS and other metadata standards;

2) How different organizations implement METS within their own repositories;

3) The nature of conformance with METS.

1. REFERENCES

[1] METS Editorial Board, version 1.6 2010. Accessed 2023. METS Primer located at http://www.loc.gov/standards/mets/METSPrimer.pdf

[2] METS website. Accessed 2023. Located at http://www.loc.gov/standards/mets/index.html

[3] COPTR Website. Accessed 2023. Located at https://coptr.digipres.org/index.php/METS_(Metadata_Encoding_and_Transmission_Standard)

[4] iPres 2022 short paper, METS version 2. Accessed 2023. Located at https://doi.org/10.17605/OSF.IO/6SEZX

[5] METS version 2 white paper. Accessed 2023. Located at https://github.com/mets/METS-schema/blob/mets2/METS2.md

[6] PREMIS website. Accessed 2023. Located at http://www.loc.gov/standards/premis/index.html

TUTORIAL: UP AND RUNNING WITH ARK PERSISTABLE IDENTIFIERS

John Kunze
Ronin Institute
USA
jakkbl@gmail.com
0000-0001-7604-8041

Donny Winston
Polyneme LLC
USA
donny@polyneme.xyz
0000-0002-8424-0604

Abstract – This half-day (3-hour) tutorial is a pragmatic introduction to the Archival Resource Key (ARK), a 22-year-old, non-paywalled identifier scheme that is widely used for persistent access to cultural and scientific information. By the end of the tutorial, attendees will know when this highly flexible scheme is appropriate to use and how to create ARKs for their respective memory organizations. No prior experience is required.

Keywords – persistent identifier, open access, URL, URI

Conference Topics – sustainably: Real and Imagined; From Theory to Practice

I. INTRODUCTION

Archival Resource Key (ARK) [1][2] identifiers help web users and content providers combat the commonly observed fragility of web addresses (URLs). The average lifetime of a URL was once said to be 100 days [3]. At the end of its life, a URL link breaks, usually giving you the dreaded "404 Not Found" error that most of us have seen. Irritating at best, it's a minor disaster for memory organizations.

In some ways ARKs are like DOIs (Digital Object Identifiers), URNs (Uniform Resource Names), and Handles. They have all been in use for over 20 years, they exist in large numbers (8.2 billion ARKs, 250 million DOIs, etc.), they are all repaired by vigilant updating of URL redirects, they support research and scholarship, and they appear in such places as the Data Citation Index, Wikipedia, and ORCiD.org profiles.

In contrast, ARKs come with no fees, no limits on how many you can create, and no metadata requirements. From the outset, ARKs were designed to be decentralized and to identify any kind of thing, whether digital, physical, or abstract.

II. TUTORIAL FORMAT

This 3-hour event (two 90-minute blocks) will be in-person or, depending on demand, remote.

The tutorial is aimed for learners who directly or indirectly manage workflows of objects used in digital libraries. The target audience includes people engaged in creating, publishing and processing web-referenceable objects from any domain in the sciences, humanities, education, law, etc. Typical attendees will be affiliated with museums, archives, libraries, data centers, and government agencies. No prior experience is required.

III. OBJECTIVES AND TOPICS

Attendees will learn when this highly flexible scheme is appropriate to use and how to create ARKs within their respective memory organizations. The following topics will be covered.

- Why ARKs – non-paywalled, decentralized, and flexible
- Use cases – Smithsonian, French National Library, Internet Archive
- Metadata for early and ongoing object development
- How to get started – one form to fill out
- Minting and assigning ARK identifiers
- Resolvers, resolution, redirection
- Object types – digital, physical, conceptual
- Persistence considerations
- Available tools

1. REFERENCES

[1] The ARK Alliance. https://arks.org/.

[2] Kunze, John. 2003. *Towards Electronic Persistence Using ARK Identifiers*. https://n2t.net/ark:/13030/c7n00zt1z.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

[3] Taylor, Nicholas. November 2011. *The Average Lifespan of a Webpage*. https://blogs.loc.gov/thesignal/2011/11/the-average-lifespan-of-a-webpage/.

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

How to Preserve Research Datasets

LABDRIVE Tutorial

Antonio Guillermo Martinez
LIBNOVA SL
Spain
a.guillermo@libnova.com

Maria Fuertes
LIBNOVA SL
Spain
mfuertes@libnova.com

Abstract – LABDRIVE is a Research Data Management and Digital Preservation platform that allows organizations to capture the research data they produce, helping them to properly manage, preserve and allow access to it, during the whole research data lifecycle, unifying and simplifying their research data management strategies.

The purpose of this tutorial is to introduce the design principles of LABDRIVE as well as explain how it works through a tutorial (a guided demonstration).

Keywords – Research Data Management, Digital Preservation, Software

Conference Topics – Immersive information, From theory to practice

I. Introduction

The EU-funded Archiver project [1] initiated a market consultation project looking for Research Data Management platforms capable to scale to the 100's of PBs in 2019. The conclusion of the market consultation was that there were neither viable nor cost-efficient platforms in the market at the time.

With the objective of helping software/platform providers to meet the need and after requesting approx. 6M€ of EU funding, a consortium led by the CERN (European Center for Nuclear Research), EMBL (European Molecular Biology Laboratory), PIC (Port d'Informació Científica – MAGIC Radio telescopes) and DESY (Deutsches Elektronen-Synchrotron) created a set of large scale data sets and use cases and initiated a Pre-Commercial Procurement (PCP) approach to competitively procure R&D services from firms in three stages, covering design, prototyping and pilot over the following 3 years.

LIBNOVA has been one of the winners over all three phases of the project (design, protype and pilot) [2], producing the LABDRIVE platform as the project result. LABDRIVE is a Research Data Management platform that supports organizations in their data management endeavors [3].

During the Archiver project, LABDRIVE has been tested and confirmed to work with High Energy physics, Astrophysics, Life Sciences and other types of large datasets (millions of files and tens of PBs) against 176 combinations of use cases, volume tests, researcher needs and organization requirements, confirming suitability and scalability of the platform for multiple Research Data Management use cases and needs.

LABDRIVE is cloud native, allowing Organizations to leverage the public/private cloud adoption if this is an objective. If not, the platform can also be deployed on premises or as hybrid cloud/on premises scenarios.

While the LIBNOVA LABDRIVE platform has been re-architected for massive scalability and specific Research Data Management use cases during the Archiver project, LIBNOVA has been the community's trusted partner for digital preservation and data management for several years. Organizations like Stanford University (HILA), Princeton University, Oxford University, The British Library, Pennsylvania

iPRES 2023: The 19th International Conference on Digital Preservation,
Champaign-Urbana, IL, US.

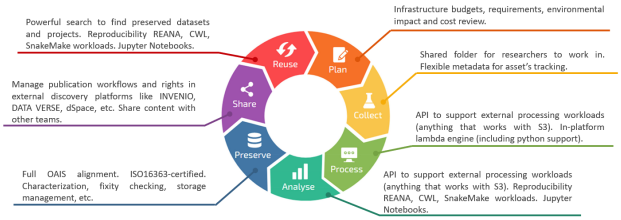
State University, Bayer and many other organizations in 17 countries are already LIBNOVA customers.

II. THE LABDRIVE PLATFORM

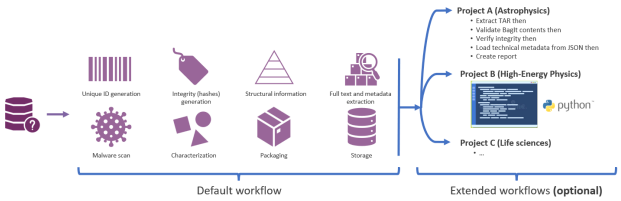
LABDRIVE is a Research Data Management and Preservation platform. It allows organizations to capture the research data they produce, helping them to properly manage, preserve and allow access to it, during the whole data lifecycle.

1. *Design principles*

1) LABDRIVE provides support over the whole data lifecycle: It allows organizations to capture the research data they produce at the initial stages of the project ("shared folder"), enabling them to properly manage, preserve, reuse and allow access to it.



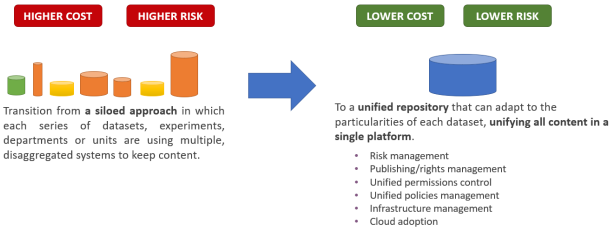
2) LABDRIVE works with many research disciplines and content types: It includes a default processing workflow, but it can be extended –using python- to support any other use case. Metadata schemas, data structures, permissions, storage, etc. can also be defined per project, so it can be adapted to multiple scenarios.



3) LABDRIVE is fully aligned with most relevant and open standards: Fully aligned to the FAIR and TRUST principles [4]. Fully conformant with OAIS [5] and fully aligned with the ISO 16363 [6]. Likewise, ISO 27001, ISO 27017 and ISO 27018-certified. GDPR compliant.

4) LABDRIVE equally supports power users and simplified use cases: Every action in the platform can be carried out using the easy-to-use web browser interface or the 300-ish Open API methods and 80+ CLI tools available.

5) As a result, LABDRIVE allows organizations to organize, unify and simplify their research data management strategies, transitioning from a siloed approach to a unified and cohesive platform, obtaining lower risks and lower costs back:



TUTORIAL CONTENT

The contents would be divided into 3 blocks and would be roughly as follows:

1. LABDRIVE Introduction
2. How it works: LABDRIVE Configuration
3. How to preserve research data: LABDRIVE Operations

2. REFERENCES

[1] ARCHIVER Project https://www.archiver-project.eu/

[2] ARCHIVER PROJECT | PILOT PHASE AWARD - THE TWO WINNERS https://archiver-project.eu/pilot-phase-award

[3] EOSC Marketplace - LIBNOVA LABDRIVE: The Ultimate Research Data Management and Digital Preservation Platform https://marketplace.eosc-portal.eu/services/libnova-labdrive-the-ultimate-research-data-management-and-digital-preservation-platform

[4] LABDRIVE support for FAIRness https://docs.libnova.com/labdrive/concepts/oais-and-iso-16363/labdrive-support-for-fairness

[5] LABDRIVE support for OAIS Conformance https://docs.libnova.com/labdrive/concepts/oais-and-iso-16363/labdrive-support-for-oais-conformance

[6] LABDRIVE - ISO 16363 certification guide https://docs.libnova.com/labdrive/concepts/oais-and-iso-16363/iso-16363-certification-guide